This CVPR workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

Deep Learning Whole Body Point Cloud Scans from a Single Depth Map

Nolan Lunscher
University of Waterloo
200 University Ave W.
nlunscher@uwaterloo.ca

John Zelek
University of Waterloo
200 University Ave W.
jzelek@uwaterloo.ca

Abstract

Personalized knowledge about body shape has numerous applications in fashion and clothing, as well as in health monitoring. Whole body 3D scanning presents a relatively simple mechanism for individuals to obtain this information about themselves without needing much knowledge of anthropometry. With current implementations however, scanning devices are large, complex and expensive. In order to make such systems as accessible and widespread as possible, it is necessary to simplify the process and reduce their hardware requirements. Deep learning models have emerged as the leading method of tackling visual tasks, including various aspects of 3D reconstruction. In this paper we demonstrate that by leveraging deep learning it is possible to create very simple whole body scanners that only require a single input depth map to operate. We show that our presented model is able to produce whole body point clouds with an accuracy of 5.19 mm.

1. Introduction

Anthropomorphic body shape is complex and is comprised of many components not easily characterized or measured. Body measurements are very important in the clothing and fashion industries, especially with the rise of online shopping where customers cannot easily try on items before purchase. This issue is especially prominent in footwear, where fit is closely tied to performance and comfort. Works towards a virtual change room hope to address these problems [33], such that a person can more conveniently try on items of clothing. With clothing, how well the items fit a person can increase their comfort as well as confidence and social well being [14]. This can be especially true for more expensive items such as suits and dresses, where a tailor is often employed to ensure a correct fit.

Generally with anthropomorphic measurements, the task of measuring is complex and requires some skill to perform accurately [10]. These difficulties in measurements can be avoided through the use of 3D scanning. Scanning can capture the complete 3D structure of an object or person without the need for the machine or operator to necessarily be an expert at how to take every measurement. Scanning also has the benefit that all possible measurements are captured, rather than only measurements at specific points [27].

Beyond clothing, understanding body measurements and shape can provide details about general health and fitness. In recent years, products such as Naked[1] and ShapeScale[2] have begun to be released with the intention of capturing 3D information about a person, and calculating various measures such as body fat percentage and muscle gains. The technologies can then be used to monitor how your body changes overtime, and offer detailed fitness tracking.

Depth map cameras and RGBD cameras, such as the XBox Kinect or Intel Realsense, have become increasingly popular in 3D scanning in recent years, and have even made their way onto mobile phones such as the iPhone X. These cameras are often favored over other forms of 3D imaging due to their ability to capture fast and accurate depth maps, even on textureless surfaces. In traditional 3D scanning methods, fully scanning an object typically requires either a moving camera or a camera array. This is because the 3D reconstruction algorithm needs to capture information from every aspect of an object in order to have knowledge of the overall shape. When scanning a non static object such as a person, neither of these solutions is ideal. With a moving camera the scanning process can take substantial time, which can allow for the person to move and break the scan [15]. In the case of a camera array, a fairly large apparatus is required, and in order to capture overlapping information for use in triangulation algorithms, cameras cannot be spaced out too sparsely, thus requiring many cameras in the array. With the camera array, both the size and number of cameras required makes this solution expensive and impractical in most circumstances.

Many of the described limitations of 3D scanning techniques have to do with the need to capture every aspect of an object. While this requirement may seem reasonable,

[1] naked.fit
[2] shapescale.com

in humans we tend to be able to overcome this by leveraging our knowledge and experience from interacting with the world. For us, we have the ability to form complete mental models of objects seen from only limited perspectives. In experiments, this has been shown by our ability to perform "mental rotation", were we are able to imagine unseen views of objects [28].

In more recent years, deep learning models have become the dominant method in many number of visual tasks. Among these growing fields, is the use of learning models for shape completion. In shape completion the goal is to produce a completed shape (typically of an object) given a limited representation. The inputs in shape completion tasks are typically either sparse surfaces [7, 26, 32] or limited viewpoints [6, 26, 29, 30, 31, 34, 35].

We apply a shape completion method to the task of full body scanning. We use a deep learning model to complete a point cloud scan when provided only a single input depth map image. With this method, whole body scanners can be made substantially less expensive and less complex. Our method of point cloud completion has previously been demonstrated to work in the application of 3D foot scanning [18]. Here we further demonstrate the flexibility of this method by applying it to the application of full body scanning, and show that it can produce high quality completed point clouds for objects as large as whole bodies.

Our method represents the shape of objects being scanned without any explicit parameterization. This method of shape learning has various advantages over parameterized techniques [8, 9], primarily that it does not require that extensive work be done to develop object specific shape parameters for each new object class to be scanned. In this way, our method can more easily be adapted to numerous scanning uses. This flexibility is especially important in anthropomorphic shape, where shape is complex and difficult to measure [10].

2. Previous Works

RGBD cameras have become very popular in applications ranging from human computer interfaces, to robotics as well as 3D scanning. These cameras are able to provide fast and accurate 3D information through a depth map which is typically produced using an infrared structured light or time of flight system. In 3D scanning, one of the more widely used algorithms is Kinect Fusion [21], which uses the video from an RGBD camera to produce a 3D mesh of an object or scene. The main drawbacks to this system and similar moving camera algorithms, are that they take a long time to move through the necessary viewpoints. Systems have also been developed that use multiple RGBD cameras to scan people [5, 15] for avatar creation, however they have their own limitations. With these systems, a large apparatus is required, as well as special calibrations that require the system to be carefully controlled. Due to the way that RGBD cameras project a pattern to determine depth, multiple cameras cannot operate viewing the same surfaces simultaneously without creating artifacts [15], thus further complicating these systems.

One approach to capturing shape information from a limited input has been to parameterize the objects shape, and develop algorithms that can predict the full set of parameters from a reduced set. This technique has previously been applied to foot modeling, where a 3D foot could be reconstructed from as few as 4 input parameters [19] or a profile and plantar outline [20]. In the case of whole body modeling, a number of works have developed methods of parameterizing body scans [4, 16, 23, 25, 42]. In these models, the shape components from dense body scans are converted into a more sparse set of surface points. Techniques such as principal component analysis can be used on the reduced set of points to further compress the number of parameters describing whole body shape. Parameterized body models provide a shape prior that can be useful in various tasks including estimating body shape under clothing [24, 36, 39] or obtaining a full 3D representation of a person from limited scans [2, 3, 37]. Similarly, learning models can be trained to determine mappings between limited inputs such as images, to a set of parameters that can then be used to reconstruct a whole body by deforming a template model [8, 9]. The main drawback to these techniques is that they are dependent on a detailed parameter set for the object being scanned. In other words, these methods cannot learn shape from an arbitrary set of object models or scans, and cannot easily be adapted across applications.

In shape completion, a common technique is to represent shape using 3D voxels, and to apply deep learning with 3D convolutions to determine the missing aspects of the object. This technique can be combined with regular 2D convolutional neural networks to process input images directly [6, 31, 34], or by operating directly on an incomplete voxel representation [7, 26, 29, 32, 35]. In either case however, the computational complexity of the 3D convolutional operations tends to limit the reconstruction resolution. In most works, the full object shape is represented within a 32^3 or 64^3 space. While this is acceptable for understanding the structure of an object, in the case of 3D scanning where we require accurate measurements of shape, the currently achievable resolutions are not sufficient, especially at the scale of an entire human body.

Another approach to shape completion is to represent shape in 2D images, and structure the problem as one of view synthesis. In view synthesis, the goal is synthesize a novel view of an object or scene given an input of one or more images. This topic has been explored greatly as a learning problem for use with color images, where techniques such as appearance flow [22, 41] and adver-

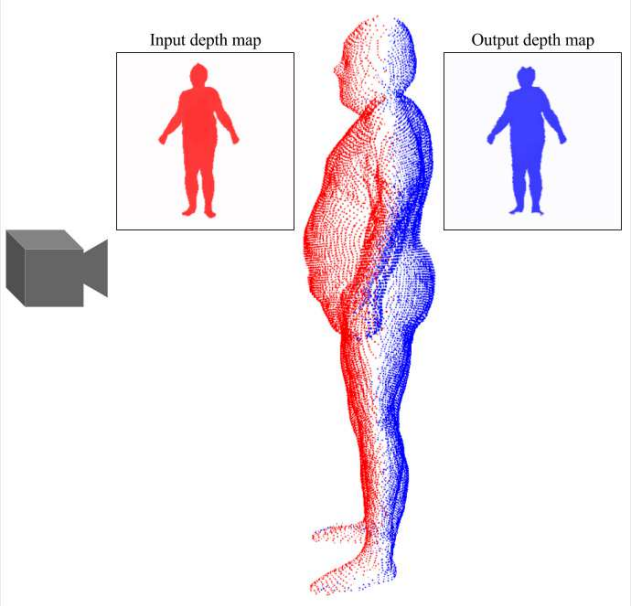

Figure 1. Depth map input and output configuration. Red: points from the input depth map, Blue: points from the output depth map.

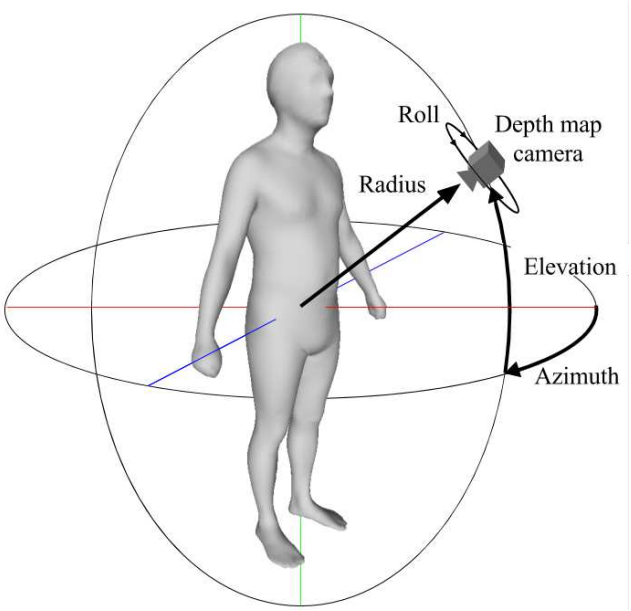

Figure 2. Depth camera pose configuration.

sarial networks [40] have been able to synthesize realistic views. In order to represent and extract shape, depth map views can be synthesized and used to form object point clouds [17, 18, 30]. Since these techniques operate on 2D images, they are able to easily take advantage of convolutional deep learning models, and allow for high resolution reconstructions with relatively low computational costs.

We follow a deep learning view synthesis approach to allow for whole body scanning from a single input depth map. This method is similar to that used for 3D foot scanning [18], with updates to allow for practical whole body scanning. Unlike typical view synthesis methods, we restrict our input to be from the front or back of the person, and take advantage of body shape to synthesize the remaining half of the person from the inverse view point. In this method, we are able to create complete point cloud scans from a single input depth map with flexibility in the exact camera pose.

3. Methods

In order to produce a completed point cloud scan from only a single input depth map, we leverage the ideas of deep learning view synthesis. Our technique is similar to that used previously for foot scanning [18], with some application specific differences. Given the popularity and availability of RGBD cameras, we structure our problem to take as input the depth map image of the front or back side of a person. Since this input depth map already contains nearly half the points required by our scan, we do not put effort into synthesizing information already on this surface. In our method, we only attempt to synthesize points that would be visible from the view on the direct opposite side from the input. The human body along this axis contains minimal self occlusions, allowing for points from these two views to be sufficient to form a complete point cloud. In order to simplify the algorithm even further, we synthesize the points on the opposite side of the body from the same camera pose as the input depth map. In this way the input and output depth maps are automatically aligned, without the need to know the exact pose the initial image was taken from. Figure 1 outlines how our input and output views are used to form a complete scan.

We restrict our input images to be taken from the front or back of the person, but we do not specify a specific camera pose defining this. We instead allow for the input viewpoint to be anywhere within a range of distances and angles in azimuth, elevation and roll as shown in Figure 2. By doing this, we do not require that the single camera scanning apparatus need to be calibrated specifically, or even mounted in any very precise way. This allows for more flexibility in how the system is laid out, and removes requirements of rigid mounting to maintain calibrations long term.

3.1. Dataset

We train a deep learning model to perform our view synthesis task using the MPII Human Shape [23] body models which are based on the CAESAR database [27]. Sample meshes from MPII Human Shape are shown in Figure 3. All mesh models are in the same standing posture, and contain 6449 vertex points with 12894 faces. We separate the 4301 body models into 80% for training and 20% for test-

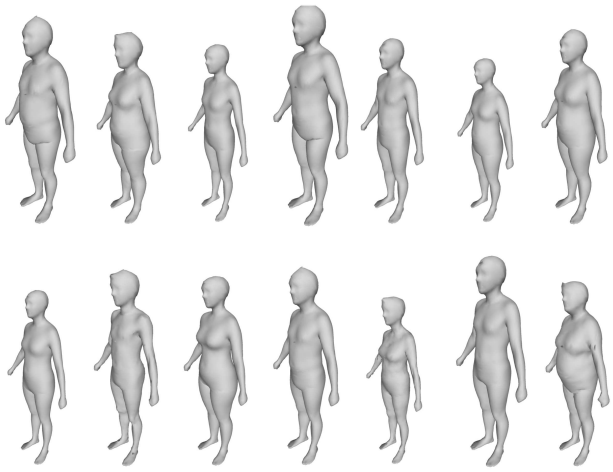

Figure 3. Sample mesh body models from MPII Human Shape [23].

Table 1. Camera pose parameter ranges for the input depth map.

Pose Parameter	Value Range	Step
Radius (m)	2.9 to 3.2	0.075
Azimuth (deg)	60 to 120 or 240 to 300	1
Elevation (deg)	-10 to 10	1
Roll (deg)	-2.4 to 2.4	0.2

ing. We use the Panda3D[3] game engine to render depth map images at a resolution of 256x256 from the range of camera poses described in Table 1. We found that depth maps of at least this resolution are required to properly represent finer components of the body such as the hands and feet.

3.2. Implementation Details

Our basic network architecture is similar to that used for foot scanning [18], with additional strided convolutional and deconvolutional [38] layers due to our higher image resolution, as well as the addition of batch normalization [12]. Our network takes in a single depth map which is passed through a set of convolutional layers followed by a set of fully connected layers, and then a set of deconvolutional layers to synthesizes an output depth map. Our network architecture is shown in Figure 4. We implemented our deep network in Tensorflow [1] on a Linux machine running an Nvidia GTX 1070 GPU. We used the Adam optimizer [13] with a mini batch size of 64 and a learning rate of 0.0001. The loss function was the mean L1 difference between the synthesized depth map and the ground truth pixels.

Complete point clouds are reconstructed by reprojecting both the input and output depth maps using the camera parameters of the scanning camera. Since the output depth map is synthesized from the same camera pose as the input, no extrinsic camera parameters are needed to form the overall point cloud. Our point clouds are additionally cleaned of outliers by 3D cropping and MATLAB's *pcdenoise* function.

[3] panda3d.org

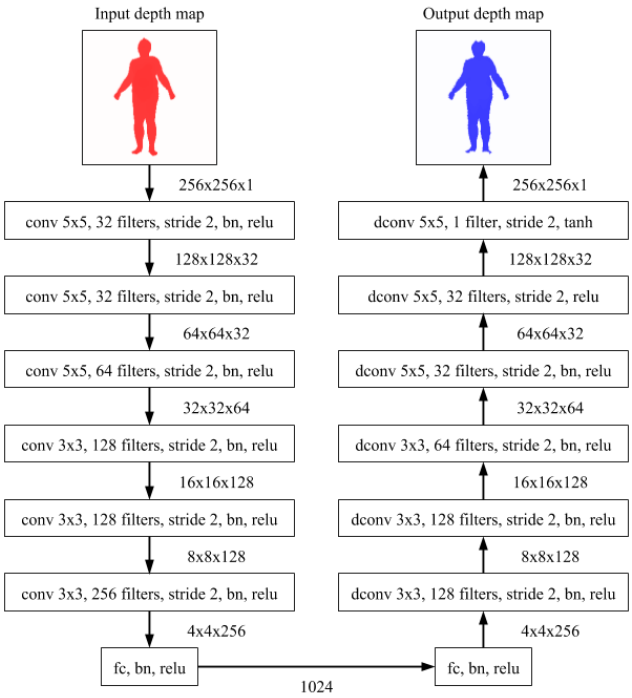

Figure 4. Network architecture.

4. Results

For each of the 860 whole body objects in our test set, we render 64 random input-output image pairs with the same camera pose parameters used in training. We trained our deep network for 1,000,000 iterations, at which point we were able to achieve a depth map L1 loss of 0.0062 on the test set. Depth map samples from our network are shown in Figure 5, along with their error distributions. As can be seen, the synthesized depth maps appear to match the ground truths very closely. Looking at the error distribution, we see that the majority of error comes from the points along the outline of the body shape. It seems as though in these regions, the network becomes unsure whether or not these pixels belong on the object surface or belong as part of the background. Generally these points will be filtered out in post-processing, however for certain features such as the feet where few points already exist, losing points to filtering can cause a larger impact on the reconstruction.

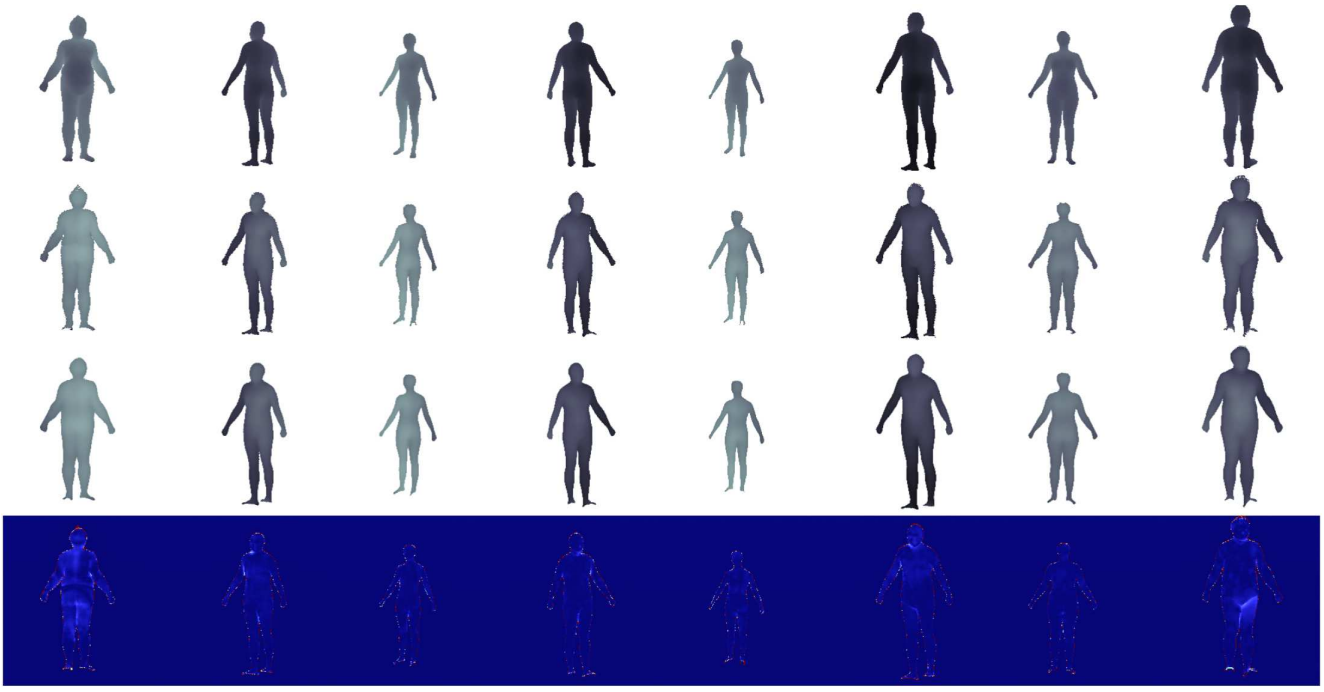

Figure 5. Synthesized depth map results. First row: input depth map, Second row: ground truth depth map, Third row: synthesized depth map, Fourth row: synthesized depth map error distribution.

4.1. Point Cloud Results

In order to quantitatively evaluate the point clouds produced by our method, we utilize a metric that has previously been used in foot scanning works [17, 18]. The metric is a two directional euclidean distance metric from a synthesized point cloud to a ground truth point cloud, where nearest neighbor is used as an approximation for point correspondence. The error $e_{syn,i}$ for each point $\mathbf{p}_{syn,i}$ in the synthesized point cloud is determined by its euclidean distance to the nearest point in the ground truth point cloud $\mathbf{p}_{gt,j}$.

$$e_{syn,i} = min_j \|\mathbf{p}_{syn,i} - \mathbf{p}_{gt,j}\|_2. \quad (1)$$

The error $e_{gt,j}$ for each point $\mathbf{p}_{gt,j}$ in the ground truth point cloud is similarly determined by its euclidean distance to the nearest point $\mathbf{p}_{syn,i}$ in the synthesized point cloud.

$$e_{gt,j} = min_i \|\mathbf{p}_{gt,j} - \mathbf{p}_{syn,i}\|_2 \quad (2)$$

These measures are normalized by the number of points in each point cloud, and then averaged together to calculate the overall point cloud error.

$$e_{total} = \frac{\frac{1}{N}\sum_i e_{syn,i} + \frac{1}{M}\sum_j e_{gt,j}}{2} \quad (3)$$

where N and M are the total number of points in the synthesized and ground truth point clouds respectively, and e_{total} is the total error for a point cloud compared to the ground truth.

Our error metric is calculated in these two directions, as it is necessary to ensure that the synthesized point cloud does not only fit a subset of the ground truth points and contain no points for the remainder. In a case such as this, far away ground truth regions would not be nearest neighbours, and would be ignored in the synthesized to ground truth point cloud error. Similarly, if the synthesized points fit the ground truth very closely but also contains additional high error points, these points would be ignored in the ground truth to synthesized point cloud error.

Using this measure, we found that our method was able to achieve an accuracy of 5.19 mm, with a standard deviation of 1.36 mm on the test set. Samples of the point clouds synthesized by our method are shown in Figure 6. As can be seen, the point cloud reconstructions match the ground truth shapes very closely for the majority of the bodys surface. Interestingly, finer details such as facial features are barely reconstructed when synthesized. These features have minimal impact in applications such as virtual change rooms, but may be a problem for other uses.

The synthesized point clouds can be observed to be more noisy around the seam between the two point clouds. This relates back to the error distribution seen in Figure 5, where the outline points are the least reliable. It can also be seen that a small gap does exist along the seam in some cases as well. This has to do both with the issue previously men-

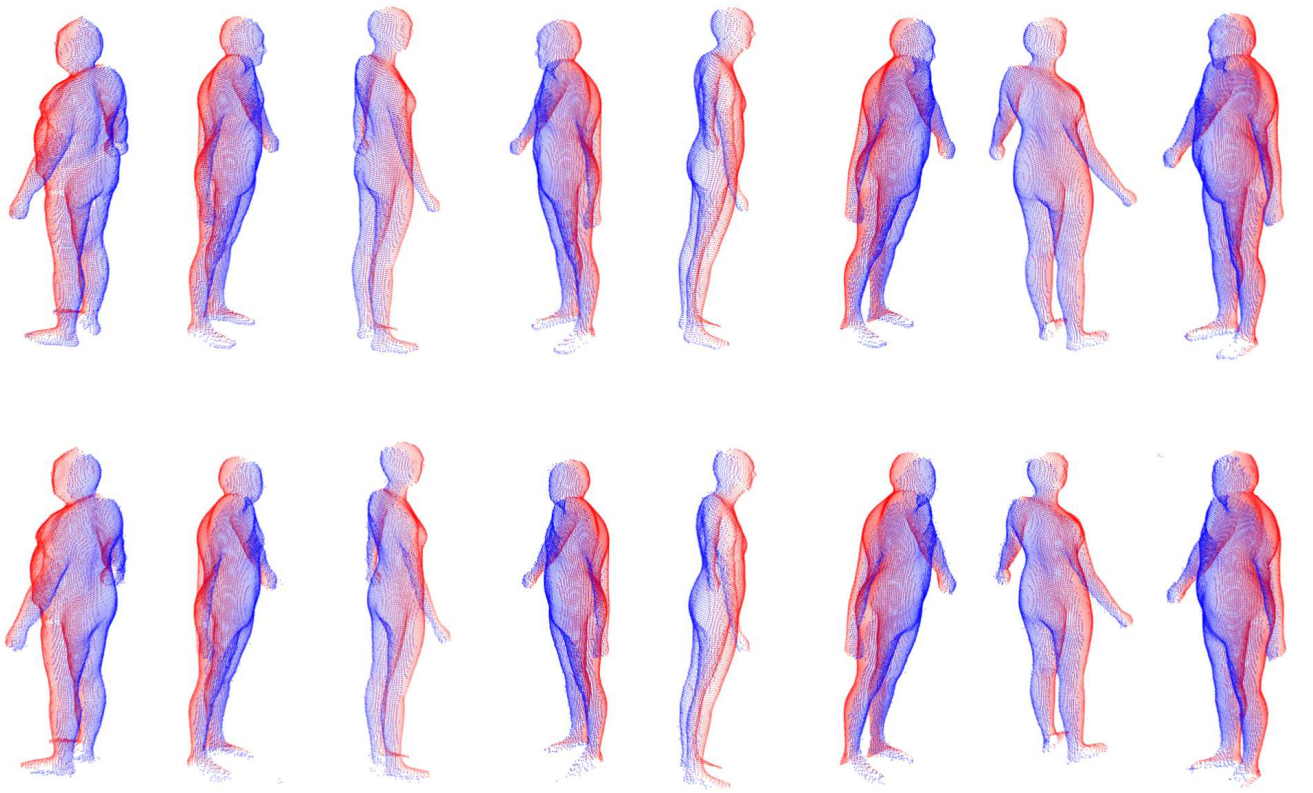

Figure 6. Completed point cloud results. Point clouds shown are reconstructions of the same depth maps in Figure 5. First row: ground truth, Second row: synthesized point cloud. Red: input depth map points, Blue: synthesized/ground truth depth map points.

tioned as well as the fact that these surfaces are perpendicular to the camera viewpoint, and thus neither depth map can capture points on this surface. Despite this gap however, the two point clouds are aligned correctly, and thus when combined still accurately represents body shape and should be sufficient for tasks of virtual try on and health monitoring. If required, 3D scanning post processing techniques exist that can be used to fill in the gaps [5].

5. Discussions and Conclusions

We have shown that our deep learning method can achieve accurate whole body scanning from a single input depth map. Our network was trained strictly from body shapes, and given no explicit shape parameters. We found that our network was able to produce point clouds with accuracies of 5.19 mm. At the scale of a whole human body, this error of only a few millimeters is rather minimal. In some applications where clothing is fitted more precisely, such as with gloves or shoes, additional scans at a closer distance or higher resolution may be required, but for general clothing, this accuracy would be sufficient in most cases.

It is important to acknowledge that our training and evaluations were conducted using MPII human shape data, which are still only approximate models of a true body scan. These body models are excellent for use in demonstrating the concept of our method and suggesting its reconstruction capabilities, however these results may differ slightly if applied to real world data. In order to apply our method on real world data, training on such data would most likely be necessary, either from scratch or as a fine tuning step.

We have also restricted the body pose to a standing posture. Our method should be flexible enough to handle more poses if trained for such cases, provided that those poses do not extensively self occlude. Perhaps shape deformations such as those used in 3D shape tracking [11] could be used to handle varying body poses, however it would require some modifications to operate from only a single input view.

Our method of whole body scanning overcomes many of the complications and limitations of more traditional scanning methods. Only a single depth map camera is required, with no strict need for special mounting or extrinsic calibration. A single camera allows for the entire scanning system to be small and affordable compared to moving camera or camera array systems. The single camera also allows for a fast scan capture time, mitigating issues where the subject has the ability to move during the scanning process. Ul-

timately, our method enables far more widespread use of 3D body scanning for applications of virtual fashion fitting, health monitoring and even avatar creation. With the simplicity of our method, it may even be useful in providing robots with 3D information of the people around them, such that they can navigate and manipulate safely around the 3D humans.

Our method has many advantages over traditional scanning methods, however it also has its share of drawbacks. The largest drawback to the system, is that it can only predict the overall shape of a persons body, rather than provide true shape information. For the majority of people the system should work sufficiently, however if someone for example has a unique shape on the surface not seen from the input view, the system will likely fail to reconstruct it. We have shown that our two view reconstruction method is flexible enough to be adapted from foot scanning [18] to whole body scanning, however its uses are still limited to shapes with minimal self occlusions along some axis. This technique may not be sufficient for use scanning shapes with more complex geometry.

In future works, we plan to investigate changes in network architecture, as well as methods of pre-processing and post-processing that may further improve our results. We also plan to explore the use of color camera images as input, as they are far more accessible than even RGBD cameras, and often are available with far higher resolutions.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. 4

[2] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. *arXiv preprint arXiv:1803.04758*, 2018. 2

[3] F. Bogo, M. J. Black, M. Loper, and J. Romero. Detailed full-body reconstructions of moving people from monocular rgb-d sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2300–2308, 2015. 2

[4] F. Bogo, J. Romero, M. Loper, and M. J. Black. Faust: Dataset and evaluation for 3d mesh registration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3794–3801, 2014. 2

[5] Y. Chen, G. Dang, Z.-Q. Cheng, and K. Xu. Fast capture of personalized avatar using two kinects. *Journal of Manufacturing Systems*, 33(1):233–240, 2014. 2, 6

[6] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European Conference on Computer Vision*, pages 628–644. Springer, 2016. 2

[7] A. Dai, C. R. Qi, and M. Nießner. Shape completion using 3d-encoder-predictor cnns and shape synthesis. *arXiv preprint arXiv:1612.00101*, 2016. 2

[8] E. Dibra, H. Jain, C. Öztireli, R. Ziegler, and M. Gross. Hs-nets: Estimating human body shape from silhouettes with convolutional neural networks. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 108–117. IEEE, 2016. 2

[9] E. Dibra, C. Öztireli, R. Ziegler, and M. Gross. Shape from selfies: Human body shape estimation using cca regression forests. In *European Conference on Computer Vision*, pages 88–104. Springer, 2016. 2

[10] M. R. Hawes and D. Sovak. Quantitative morphology of the human foot in a north american population. *Ergonomics*, 37(7):1213–1226, 1994. 1, 2

[11] C.-H. Huang, B. Allain, E. Boyer, J.-S. Franco, F. Tombari, N. Navab, and S. Ilic. Tracking-by-detection of 3d human shapes: from surfaces to volumes. *IEEE transactions on pattern analysis and machine intelligence*, 2017. 6

[12] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015. 4

[13] D. P. Kingma and J. L. Ba. Adam: a Method for Stochastic Optimization. In *International Conference on Learning Representations 2015*, pages 1–15, 2015. 4

[14] T. R. Kinley. Fit and shopping preferences by clothing benefits sought. *Journal of Fashion Marketing and Management: An International Journal*, 14(3):397–411, 2010. 1

[15] S. Lin, Y. Chen, Y.-K. Lai, R. R. Martin, and Z.-Q. Cheng. Fast capture of textured full-body avatar with rgb-d cameras. *The Visual Computer*, 32(6-8):681–691, 2016. 1, 2

[16] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):248, 2015. 2

[17] N. Lunscher and J. Zelek. Deep learning anthropomorphic 3d point clouds from a single depth map camera viewpoint. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, 2017. 3, 5

[18] N. Lunscher and J. Zelek. Point cloud completion of foot shape from a single depth map for fit matching using deep learning view synthesis. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, 2017. 2, 3, 4, 5, 7

[19] A. Luximon and R. S. Goonetilleke. Foot shape modeling. *Human Factors*, 46(2):304–315, 2004. 2

[20] A. Luximon, R. S. Goonetilleke, and M. Zhang. 3d foot shape generation from 2d information. *Ergonomics*, 48(6):625–641, 2005. 2

[21] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on*, pages 127–136. IEEE, 2011. 2

[22] E. Park, J. Yang, E. Yumer, D. Ceylan, and A. C. Berg. Transformation-grounded image generation network for novel 3d view synthesis. *arXiv preprint arXiv:1703.02921*, 2017. 2

[23] L. Pishchulin, S. Wuhrer, T. Helten, C. Theobalt, and B. Schiele. Building statistical shape spaces for 3d human modeling. *CoRR*, abs/1503.05860, 2015. 2, 3, 4

[24] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics (TOG)*, 36(4):73, 2017. 2

[25] G. Pons-Moll, J. Romero, N. Mahmood, and M. J. Black. Dyna: A model of dynamic human shape in motion. *ACM Transactions on Graphics (TOG)*, 34(4):120, 2015. 2

[26] D. J. Rezende, S. A. Eslami, S. Mohamed, P. Battaglia, M. Jaderberg, and N. Heess. Unsupervised learning of 3d structure from images. In *Advances in Neural Information Processing Systems*, pages 4996–5004, 2016. 2

[27] K. M. Robinette, H. Daanen, and E. Paquet. The caesar project: a 3-d surface anthropometry survey. In *3-D Digital Imaging and Modeling, 1999. Proceedings. Second International Conference on*, pages 380–386. IEEE, 1999. 1, 3

[28] R. N. Shepard and J. Metzler. Mental rotation of three-dimensional objects. *Science*, 171(3972):701–703, 1971. 2

[29] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser. Semantic scene completion from a single depth image. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[30] M. Tatarchenko, A. Dosovitskiy, and T. Brox. Multi-view 3d models from single images with a convolutional network. In *European Conference on Computer Vision*, pages 322–337. Springer, 2016. 2, 3

[31] J. Wu, Y. Wang, T. Xue, X. Sun, W. T. Freeman, and J. B. Tenenbaum. MarrNet: 3D Shape Reconstruction via 2.5D Sketches. In *Advances In Neural Information Processing Systems*, 2017. 2

[32] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1912–1920, 2015. 2

[33] S. Wuhrer, L. Pishchulin, A. Brunton, C. Shu, and J. Lang. Estimation of human body shape and posture under clothing. *Computer Vision and Image Understanding*, 127:31–42, 2014. 1

[34] X. Yan, J. Yang, E. Yumer, Y. Guo, and H. Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *Advances in Neural Information Processing Systems*, pages 1696–1704, 2016. 2

[35] B. Yang, H. Wen, S. Wang, R. Clark, A. Markham, and N. Trigoni. 3d object reconstruction from a single depth view with adversarial learning. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2

[36] J. Yang, J.-S. Franco, F. Hétroy-Wheeler, and S. Wuhrer. Estimation of human body shape in motion with wide clothing. In *European Conference on Computer Vision*, pages 439–454. Springer, 2016. 2

[37] M. Ye, H. Wang, N. Deng, X. Yang, and R. Yang. Real-time human pose and shape estimation for virtual try-on using a single commodity depth camera. *IEEE transactions on visualization and computer graphics*, 20(4):550–559, 2014. 2

[38] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus. Deconvolutional networks. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2528–2535. IEEE, 2010. 4

[39] C. Zhang, S. Pujades, M. Black, and G. Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2017. 2

[40] B. Zhao, X. Wu, Z.-Q. Cheng, H. Liu, and J. Feng. Multi-view image generation from a single-view. *arXiv preprint arXiv:1704.04886*, 2017. 3

[41] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros. View synthesis by appearance flow. In *European Conference on Computer Vision*, pages 286–301. Springer, 2016. 2

[42] S. Zuffi and M. J. Black. The stitched puppet: A graphical model of 3d human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3537–3546, 2015. 2