DSPP: Deep Shape and Pose Priors of Humans

Shanfeng Hu
shanfeng.hu@northumbria.ac.uk
Department of Computer and Information Sciences
Northumbria University
Newcastle upon Tyne, UK

Hubert P. H. Shum*
hubert.shum@northumbria.ac.uk
Department of Computer and Information Sciences
Northumbria University
Newcastle upon Tyne, UK

Antonio Mucherino
antonio.mucherino@irisa.fr
IRISA
University of Rennes 1
Rennes, France

ABSTRACT

The prior knowledge of real human body shapes and poses is fundamental in computer games and animation (e.g. performance capture). Linear subspaces such as the popular SMPL model have a limited capacity to represent the large geometric variations of human shapes and poses. What is worse is that random sampling from them often produces non-realistic humans because the distribution of real humans is more likely to concentrate on a non-linear manifold instead of the full subspace. Towards this problem, we propose to learn human shape and pose manifolds using a more powerful deep generator network, which is trained to produce samples that cannot be distinguished from real humans by a deep discriminator network. In contrast to previous work that learn both the generator and discriminator in the original geometry spaces, we learn them in the more representative latent spaces discovered by a shape and a pose auto-encoder network respectively. Random sampling from our priors produces higher-quality human shapes and poses. The capacity of our priors is best applied to applications such as virtual human synthesis in games.

CCS CONCEPTS

• **Computing methodologies → Shape modelling**.

KEYWORDS

human shape modelling, human pose modelling, generative adversarial networks, deep learning

ACM Reference Format:
Shanfeng Hu, Hubert P. H. Shum, and Antonio Mucherino. 2019. DSPP: Deep Shape and Pose Priors of Humans. In *Motion, Interaction and Games (MIG '19), October 28–30, 2019, Newcastle upon Tyne, United Kingdom.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3359566.3360051

1 INTRODUCTION

3D human body shapes and poses are ubiquitous in computer games and animations. Modelling their distributions is a fundamental building block for automatically synthesising realistic-looking human characters and animations.

The drawback of linear subspace methods such as the SMPL model [Loper et al. 2015] is that random sampling far from the centre (i.e. the average human shape and pose) of the resulting Gaussian distributions would produce non-realistic humans [Kanazawa et al. 2018]. This is because the distributions of real human shapes and poses are more likely to be locally supported on non-linear manifolds, instead of on the full subspaces where the Gaussian distributions are globally supported on.

The more recent work on generative adversarial networks (GANs) for non-linear distribution modelling [Goodfellow et al. 2014] still cannot produce satisfactory results, because they embed the shape and pose manifolds directly in the high-dimensional geometry spaces [Kanazawa et al. 2018]. The training can be difficult because measuring the overlap of two manifolds (i.e. the real and the generated) in high-dimensional spaces is difficult.

In this paper, we propose to learn the distributions of real human shapes and poses using two separate GANs, not in the original high-dimensional geometry spaces but in the more representative low-dimensional latent spaces discovered by a shape and a pose auto-encoder network respectively. The motivation is that the dimensions of the real shape and pose manifolds should be independent of the ambient spaces. As a result, measuring their overlaps with the generated shape and pose manifolds can be made easier in a much lower-dimensional ambient space. Therefore, we train a shape encoder (similarly a pose encoder) to embed real human shapes into a low-dimensional hidden space, while training a shape decoder (similarly a pose decoder) to reconstruct the input. Jointly, we train a shape generator (similarly a pose generator) to transform a standard Gaussian distribution into this space, in which a shape discriminator (similarly a pose discriminator) is also trained to approximate the distribution distance for the generator to minimise.

We propose the two following contributions in this paper:

- We propose to learn the non-linear manifolds of real human shapes and poses in the low-dimensional auto-encoding spaces, rather than in the original high-dimensional geometry spaces.
- We demonstrate the capacity of our learned priors by generating high-quality human shapes and poses via random sampling. We release our source code[1] to facilitate the synthesis of realistic humans in real-time (around 5ms using a GTX 1080 graphics card).

*Correspondence author

MIG '19, October 28–30, 2019, Newcastle upon Tyne, United Kingdom

ACM ISBN 978-1-4503-6994-7/19/10…$15.00
https://doi.org/10.1145/3359566.3360051

[1] https://drive.google.com/open?id=1y-aPe8FGzetxnY3FpSESci3U59KgQSUJ

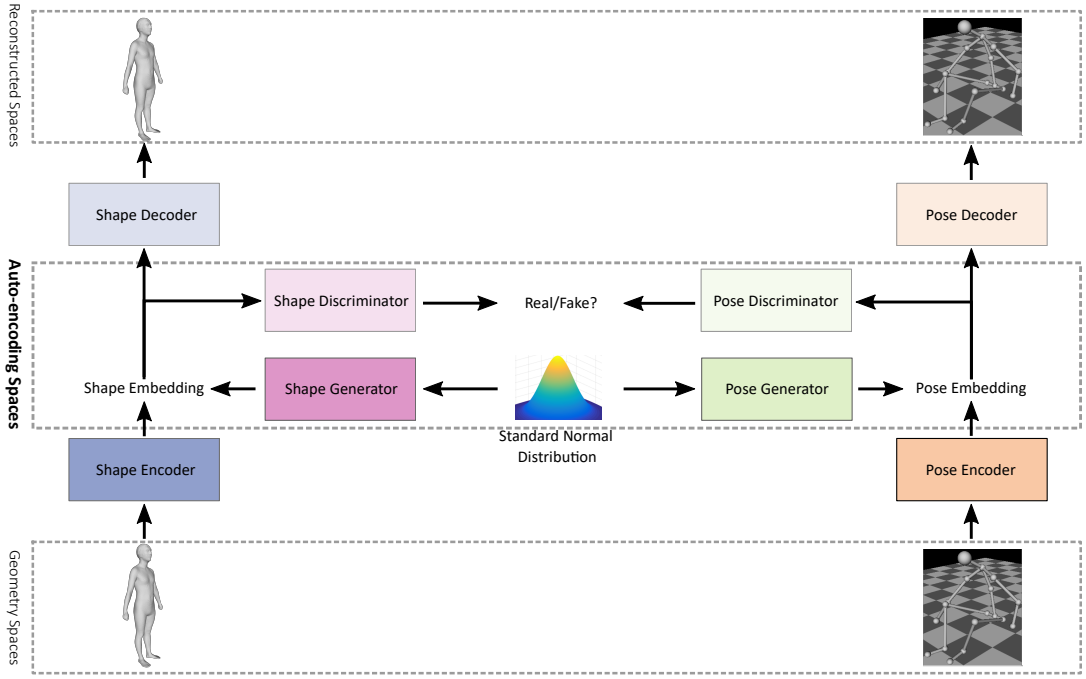

Figure 1: We jointly train a pair of shape encoder/decoder and a pair of pose encoder/decoder so that we can learn the non-linear manifolds of human shape and pose in the low-dimensional auto-encoding ambient spaces. Synthesising realistic human shapes and poses amounts to sampling from a standard normal distribution, then applying the corresponding generators, and finally applying the corresponding decoders.

2 RELATED WORK

Here, we briefly discuss existing work on modelling the prior distributions of human body shapes and poses. Because human shapes and poses have geometric regularities, they are commonly assumed to be lying on the underlying low-dimensional manifolds, which are embedded in the original geometry ambient spaces.

Linear subspace methods assume the shape and pose manifolds to be linearly embedded in the geometry spaces [Allen et al. 2003; Anguelov et al. 2005; Blanz et al. 1999; Loper et al. 2015], which can be computed using the linear PCA dimension reduction method. The drawback of linear methods is that they mostly use Gaussian as the generating distributions, which are globally supported on the full subspace. Sampling in the full subspace may produce non-realistic results [Kanazawa et al. 2018].

More recent work makes a more realistic assumption of human shape and pose manifolds, that they are non-linearly embedded in the geometry spaces. The variational auto-encoding (VAE) framework of [Kingma and Welling 2013] has been exploited to learn human shape and pose manifolds in [Tan et al. 2018] and [Habibie et al. 2017] respectively. There are also methods using GANs [Goodfellow et al. 2014] to model human shape and pose manifolds [Chen et al. 2017; Gokaslan et al. 2018; Kanazawa et al. 2018]. We prefer GANs over VAE for non-linear distribution modelling in this work because the former have been extensively validated to be capable of producing sharp samples (e.g. images and shapes), while the latter tend to produce over-smoothed results. Our work is based on that of [Makhzani et al. 2015] that combines auto-encoders with GANs for more effective distribution modelling. The main contribution of this work is our finding that learning the shape and pose manifolds in the auto-encoding spaces with GANs produces high-quality samples.

3 DSPP: DEEP SHAPE AND POSE PRIORS OF HUMANS

We focus on modelling the probability distributions of real human body shapes $x \sim \mathbf{p}(x)$ and real human body poses $y \sim \mathbf{p}(y)$. In computer gaming and animation, it is typical to represent a human shape $x \in \mathbb{R}^{N\times 3}$ as a 3D point cloud with a given mesh topology, and a human pose $y \in \mathbb{R}^{M\times 3}$ as an array of 3D joint Euler angles. Combining the two using skinning techniques can produce a deformed human shape in the given pose [Wang et al. 2015].

The challenge of modelling $\mathbf{p}(x)$ and $\mathbf{p}(y)$ is that they are high-dimensional distributions but can only be specified as the empirical distributions of scanned human shapes and motion-captured human poses in practice. That is, sampling from them is equivalent to sampling from a shape dataset and a pose dataset respectively. As a result, our task is approximating them using continuous distributions $\hat{\mathbf{p}}(x)$ and $\hat{\mathbf{p}}(y)$ so that the drawn samples are visually similar to that from the given datasets.

The overview of our method is illustrated in Fig. 1. The key is that we assume $\hat{\mathbf{p}}(x)$ and $\hat{\mathbf{p}}(y)$ to be locally supported on two

low-dimensional manifolds respectively, whose dimensions are independent of the ambient spaces the manifolds are embedded in. Our method departs from the previous work on linear subspaces that assume both distributions to be Gaussian with a global support. It is also in contrast with the recent work on generative distribution modelling that embeds the manifolds in the original high-dimensional geometry space. Particularly, our method jointly learns some low-dimensional ambient spaces and the manifolds embedded within, which is shown to produce higher-quality samples.

3.1 Auto-encoding Ambient Spaces

To find the low-dimensional ambient spaces for generative distribution modelling, we learn two pairs of deep encoders $\{f_{\text{shape}}, f_{\text{pose}}\}$ and decoders $\{h_{\text{shape}}, h_{\text{pose}}\}$, by minimising the two following reconstruction losses respectively:

$$l(x, f_{\text{shape}}, h_{\text{shape}}) = \mathbf{E}_{x\sim\mathbf{p}(x)}\|x - h_{\text{shape}}(f_{\text{shape}}(x))\|_F^2 \quad (1)$$

$$l(y, f_{\text{pose}}, h_{\text{pose}}) = \mathbf{E}_{y\sim\mathbf{p}(y)}\|y - h_{\text{pose}}(f_{\text{pose}}(y))\|_F^2 \quad (2)$$

where the expectations of the squared Frobenius norms[2] are taken with respect to the empirical shape and pose distributions, $\mathbf{p}(x)$ and $\mathbf{p}(y)$, respectively.

We denote $z_x = f_{\text{shape}}(x)$ and $z_y = f_{\text{pose}}(y)$ as the learned hidden representations of real human shapes and poses respectively. This allows us to obtain $z_x \sim \mathbf{p}(z_x)$ and $z_y \sim \mathbf{p}(z_y)$ as the low-dimensional empirical distributions of real human shape and pose representations. Because we can use a much smaller dimension for z_x and z_y respectively, our task of approximating the original high-dimensional empirical distributions $\mathbf{p}(x)$ and $\mathbf{p}(y)$ can be greatly simplified to model $\mathbf{p}(z_x)$ and $\mathbf{p}(z_y)$ instead.

3.2 Generative Modelling in the Auto-encoding Spaces

Now, we seek to find a pair of continuous distributions $\hat{\mathbf{p}}(z_x)$ and $\hat{\mathbf{p}}(z_y)$ so that they can be made close to their empirical distribution counterparts, by minimising some distribution distance between them. Assuming the continuous distributions to be Gaussian and taking the distance as the Kullback–Leibler divergence, as done in the previous work on linear subspace methods, recover the globally supported Gaussian distributions that produce non-realistic samples far from the centres.

Instead, we assume $\hat{\mathbf{p}}(z_x)$ and $\hat{\mathbf{p}}(z_y)$ to be only locally supported on the low-dimensional human shape and pose manifolds, which are embedded in the auto-encoding shape and pose ambient spaces respectively. Therefore, we leverage the state-of-the-art GANs for discovering such non-linear manifolds [Goodfellow et al. 2014]. We train a shape generator $\varphi_{\text{shape}} : z \rightarrow z_x$ and a pose generator $\varphi_{\text{pose}} : z \rightarrow z_y$ that embed a random sample $z \sim \mathcal{N}(0, I)$ from a low-dimensional standard Gaussian distribution into the shape and pose ambient spaces respectively. This way, synthesising human shapes and poses from the embedded manifolds is equivalent to sampling z and then applying the two generator functions respectively.

To fit the embedded shape and pose manifolds to the empirical distributions $\mathbf{p}(z_x)$ and $\mathbf{p}(z_y)$, we take the popular discrimination loss:

$$l(\hat{\mathbf{p}}(z_x), \mathbf{p}(z_x)) = -\mathbf{E}_{z\sim\mathcal{N}(0,I)}[\log\psi_{\text{shape}}(\varphi_{\text{shape}}(z))] \quad (3)$$

$$l(\hat{\mathbf{p}}(z_y), \mathbf{p}(z_y)) = -\mathbf{E}_{z\sim\mathcal{N}(0,I)}[\log\psi_{\text{pose}}(\varphi_{\text{pose}}(z))] \quad (4)$$

where $\psi_{\text{shape}} : z_x \rightarrow (0, 1)$ and $\psi_{\text{pose}} : z_y \rightarrow (0, 1)$ are the separately trained discriminator functions that tell whether a given shape and a pose are real samples or not: near-1 probabilities mean real and near-0 probabilities mean synthetic. Minimising (3) and (4) amounts to finding the shape and pose generators that produce synthetic samples indistinguishable from the real ones according to the corresponding discriminators [Goodfellow et al. 2014].

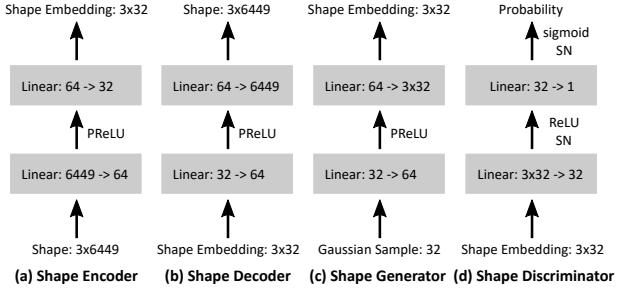

Figure 2: The neural network architectures of our proposed shape encoder, shape decoder, shape generator, and shape discriminator.

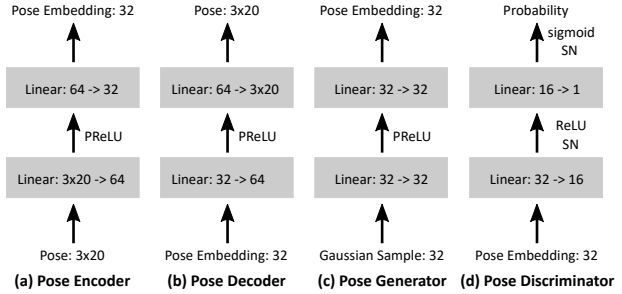

Figure 3: The neural network architectures of our proposed pose encoder, pose decoder, pose generator, and pose discriminator.

3.3 Datasets, Architectures, and Training

To validate the effectiveness of our method, we jointly train our human shape and pose distribution modelling system on the MPII Human Shape dataset [Pishchulin et al. 2017] and the SFU Motion Capture dataset [SFU 2016]. The former contains 4,308 3D human body shapes registered to a common mesh topology, with each shape consisting of 6,449 surface points. The latter provides 3D human motion capture clips covering various activities such as walking, running, dancing, and interactions. We extract over 100,000 poses from these clips using a regular sampling rate of 4 frames/second. We represent each pose using the 3D Euler angles of 20 body joints defined in [Pishchulin et al. 2017], excluding that of the Hips root joint as it describes global rotations.

[2] $\|X\|_F^2 = \text{trace}(X^T X)$

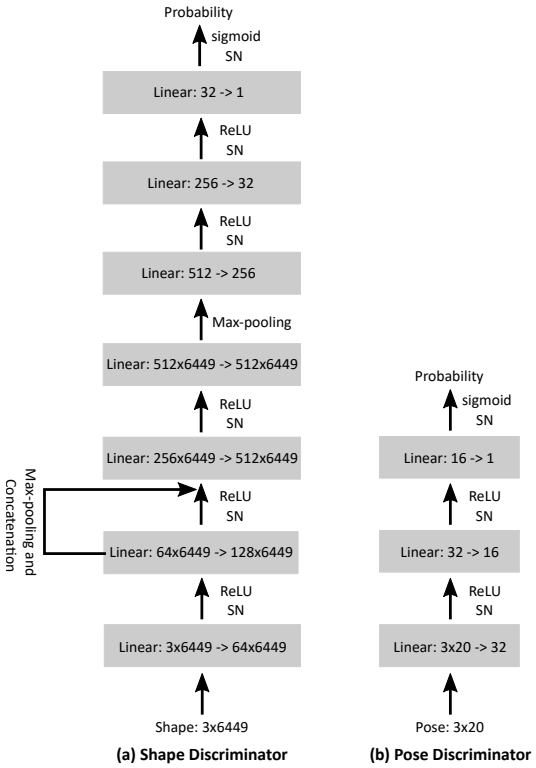

Figure 4: The architectures of the baseline shape and pose discriminators that work in the original geometry ambient spaces.

We illustrate the neural network architectures of our shape and pose subsystems in Fig. 2 and 3 respectively. We also show the baseline shape and pose discriminator architectures that work in the original geometry spaces in Fig. 4, where we use the max-pooling operator to aggregate point features for global context modelling [Qi et al. 2017]. We use the Parametric Rectified Linear Unit (PReLU) activation function for the encoders, decoders, and generators [He et al. 2015], while using the Rectified Linear Unit (ReLU) [Nair and Hinton 2010] for the discriminators. Importantly, we follow the method of [Miyato et al. 2018] to normalise the spectral norm (SN) of each linear transformation in the discriminators to be 1, which effectively stabilises the training of GANs. In each of 100,000 training cycles, we first randomly sample a pair of real shape and pose to train the corresponding encoders and decoders. We then train the discriminators using the updated embeddings and a pair of synthetic samples computed from the respective generators. Finally, we train the two generators based on the corresponding updated discriminators. We use the Adam optimiser with the default settings for training [Kingma and Ba 2014].

We implement our system in PyTorch (V1.1) on a PC with a GTX 1080 graphics card with 8GB graphics memory. Each training step takes around 70ms, and generating a pair of shape and pose takes around 5ms. Our code is publicly available from this link: https://drive.google.com/open?id=1y-aPe8FGzetxnY3FpSESci3U59KgQSUJ

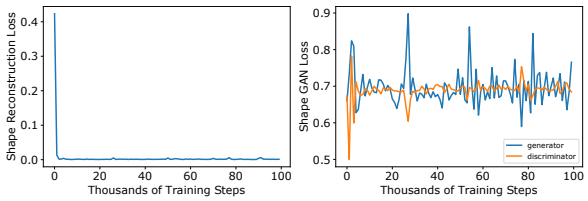

Figure 5: The loss of our shape encoder, shape decoder, shape generator, and shape discriminator.

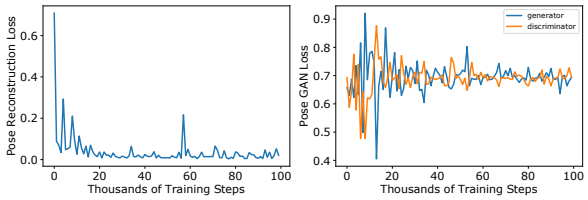

Figure 6: The loss of our pose encoder, pose decoder, pose generator, and pose discriminator.

4 RESULTS

We show the training losses of our proposed human shape and pose modelling subsystems in Fig. 5 and 6 respectively. As the training progresses, both the shape and pose reconstruction losses decrease rapidly. This shows that the shape and pose encoders effectively compress the original geometry input into the more representative spaces, from which the corresponding decoders successfully reconstruct the input. The fluctuations of the shape and pose GAN losses are desirable, which indicate that the generators are continuously learning to produce samples that cannot easily be distinguished by the corresponding discriminators.

We compare our randomly sampled human shapes with that sampled from the method of using the baseline shape discriminator (Fig. 4, left) in Fig. 7. Our synthesised shapes look both realistic and diverse, resembling closely with the samples from the ground-truth dataset [Pishchulin et al. 2017]. In comparison, the baseline method produces noticeable noises around the head, chest, and feet regions. We also compare our randomly sampled human poses with that sampled from the method of using the baseline pose discriminator (Fig. 4, right) in Fig. 8. Similarly, our generated poses look considerably more realistic than the baseline results. Together, these results validate the effectiveness of modelling shape and pose manifolds in the auto-encoding spaces.

5 CONCLUSION

We introduced the idea of modelling the non-linear manifolds of human shapes and poses in the auto-encoding ambient spaces. We discovered the spaces by learning a pair of encoder and decoder for human shapes and poses respectively. The low-dimension of such spaces allow us to more effectively learn human shape and pose manifolds, using the powerful non-linear distribution modelling GAN. The learned manifolds, as embedded in the auto-encoding ambient spaces, allow for the synthesis of realistic human shapes and poses. Our results showed that our method produces higher-quality

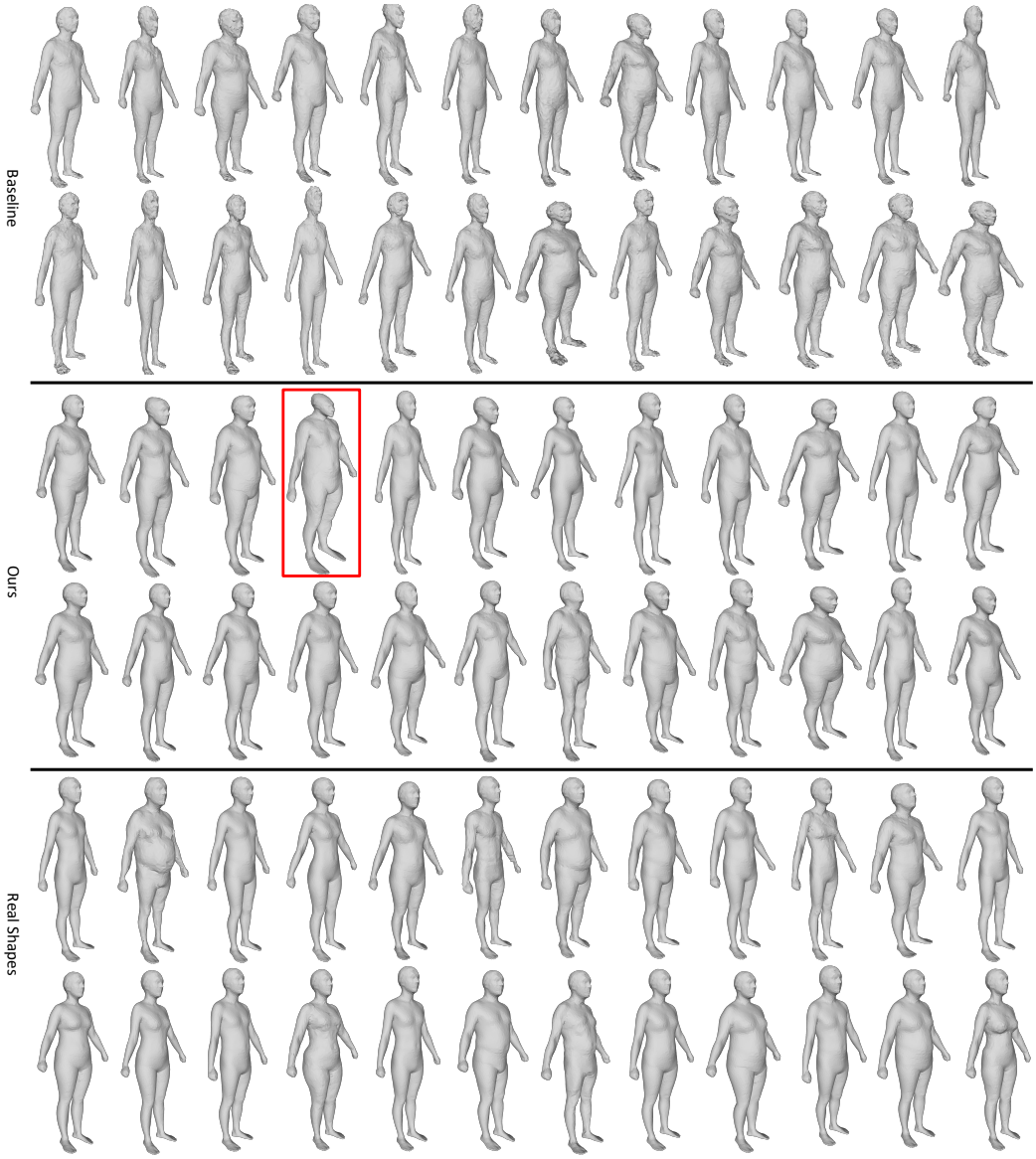

Figure 7: Comparison of our randomly sampled human shapes with that sampled from the baseline method of learning discrimination in the original shape space. We use red rectangles to indicate our samples that do not look realistic.

samples, comparing with the method of distribution modelling in the original geometry spaces.

As a future work, we are interested in learning to realistically deform the synthesised shapes using the synthesised poses. This will open the door of automatic human character synthesis in applications such as animation and gaming. Although the method of [Loper et al. 2015] permits such applications, it relies on the traditional skinning technique and could result in non-realistic deformations. The work of [Bailey et al. 2018] that automatically learns deformations is worthy exploration. We are also interested in some quantitative metrics for evaluating the realism of automatically synthesised humans.

ACKNOWLEDGMENTS

This work was supported by the Royal Society (Ref: IES\R2\181024). The authors also thank the anonymous reviewers for their constructive comments.

REFERENCES

2016. SFU Motion Capture dataset. http://mocap.cs.sfu.ca/

Brett Allen, Brian Curless, Brian Curless, and Zoran Popović. 2003. The space of human body shapes: reconstruction and parameterization from range scans. In *ACM TOG*, Vol. 22. 587–594.

Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. 2005. SCAPE: shape completion and animation of people. In *ACM TOG*, Vol. 24. 408–416.

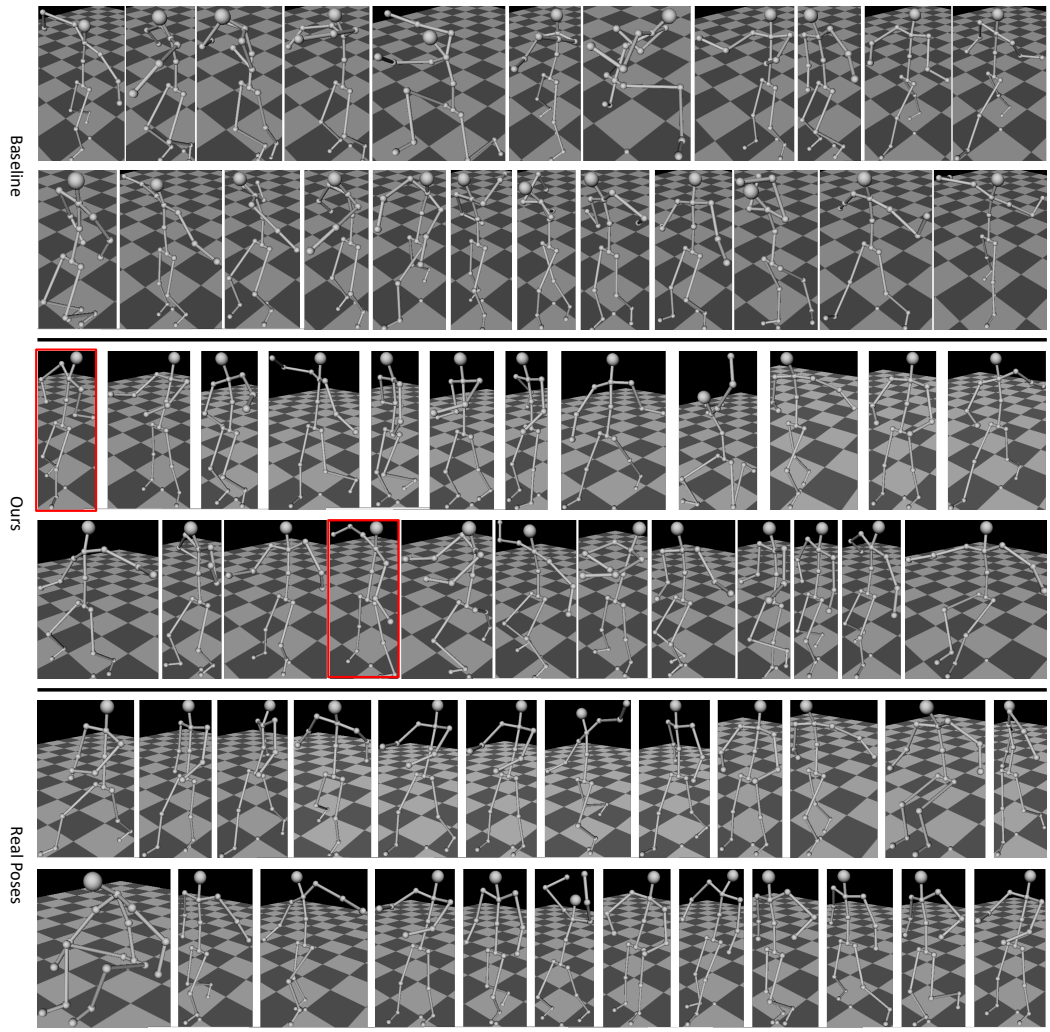

Figure 8: Comparison of our randomly sampled human poses with that sampled from the baseline method of learning discrimination in the original pose space. We use red rectangles to indicate our samples that do not look realistic.

Stephen W Bailey, Dave Otte, Paul Dilorenzo, and James F O'Brien. 2018. Fast and deep deformation approximations. *ACM TOG* 37, 4 (2018), 119.

Volker Blanz, Thomas Vetter, et al. 1999. A morphable model for the synthesis of 3D faces. 99, 1999 (1999), 187–194.

Yu Chen, Chunhua Shen, Xiu-Shen Wei, Lingqiao Liu, and Jian Yang. 2017. Adversarial posenet: A structure-aware convolutional network for human pose estimation. *ICCV* (2017), 1212–1221.

Aaron Gokaslan, Vivek Ramanujan, Daniel Ritchie, Kwang In Kim, and James Tompkin. 2018. Improving shape deformation in unsupervised image-to-image translation. *ECCV* (2018), 649–665.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *NIPS* (2014), 2672–2680.

Ikhsanul Habibie, Daniel Holden, Jonathan Schwarz, Joe Yearsley, and Taku Komura. 2017. A Recurrent Variational Autoencoder for Human Motion Synthesis. *BMVC* (2017).

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *ICCV* (2015), 1026–1034.

Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. 2018. End-to-end recovery of human shape and pose. *CVPR* (2018), 7122–7131.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2015. SMPL: A skinned multi-person linear model. *ACM TOG* 34, 6 (2015), 248.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. 2015. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644* (2015).

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. 2018. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957* (2018).

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. *ICML* (2010), 807–814.

Leonid Pishchulin, Stefanie Wuhrer, Thomas Helten, Christian Theobalt, and Bernt Schiele. 2017. Building statistical shape spaces for 3d human modeling. *Pattern Recognition* 67 (2017), 276–286.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. *CVPR* (2017), 652–660.

Qingyang Tan, Lin Gao, Yu-Kun Lai, and Shihong Xia. 2018. Variational autoencoders for deforming 3d mesh models. *CVPR* (2018), 5841–5850.

Yu Wang, Alec Jacobson, Jernej Barbič, and Ladislav Kavan. 2015. Linear subspace design for real-time shape deformation. *ACM TOG* 34, 4 (2015), 57.