Indirect deep structured learning for 3D human body shape and pose prediction

Jun Kai Vince Tan
jkvt2@cam.ac.uk
Ignas Budvytis
ib255@cam.ac.uk
Roberto Cipolla
rc10001@cam.ac.uk

Department of Engineering
University of Cambridge
Cambridge, UK

Abstract

In this paper we present a novel method for 3D human body shape and pose prediction. Our work is motivated by the need to reduce our reliance on costly-to-obtain ground truth labels. To achieve this, we propose training an encoder-decoder network using a two step procedure as follows. During the first step, a decoder is trained to predict a body silhouette using SMPL [2] (a statistical body shape model) parameters as an input. During the second step, the whole network is trained on real image and corresponding silhouette pairs while the decoder is kept fixed. Such a procedure allows for an *indirect* learning of body shape and pose parameters from real images without requiring any ground truth parameter data.

Our key contributions include: (a) a novel encoder-decoder architecture for 3D body shape and pose prediction, (b) corresponding training procedure as well as (c) quantitative and qualitative analysis of the proposed method on artificial and real image datasets.

1 Introduction

The recent advent of deep neural networks and, in particular, Convolutional Neural Networks (CNNs) has significantly increased the state-of-the-art performance in data-rich problems of computer vision such as image classification [17], object detection [3] and semantic segmentation [10]. While datasets for the aforementioned computer vision problems now frequently reach hundreds of thousands of images, datasets for problems that require complex labelling procedures tend to remain modest in size.

In this work we apply indirect learning from artificial data as an alternative method for tackling the problem of 3D human shape and pose prediction. In particular, we propose learning an encoder-decoder network which employs a two-phase training procedure as illustrated in Figure 1. In the first phase, a decoder learns to predict a human silhouette from SMPL [2] parameters. In the second phase, a full encoder-decoder network is trained to predict both human silhouette and 3D model parameters from a single RGB image. During this phase, the decoder weights and biases are fixed, allowing us to train only on real image and corresponding silhouette pairs without the hard-to-obtain corresponding 3D pose and shape parameters. We provide a quantitative and qualitative evaluation of our proposed method on

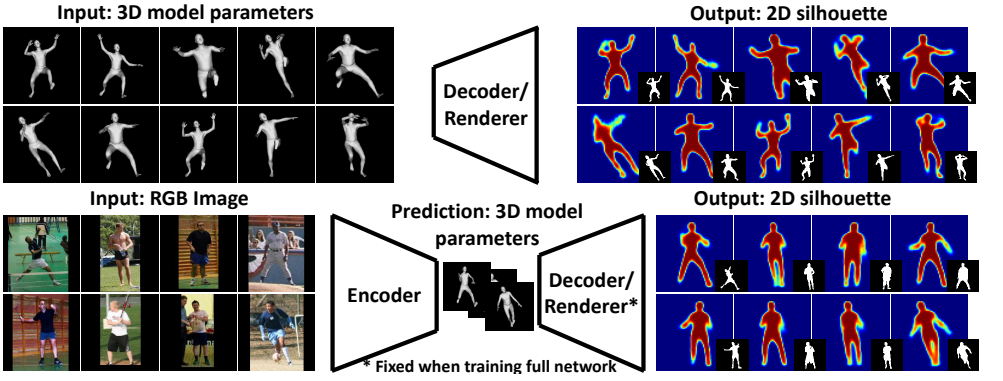

Figure 1: This figure illustrates our proposed method for indirect learning using artificial data. We first train a decoder (renderer) to predict a 2D human body silhouette from parameters of SMPL [7] - a statistical human body model. We then use pairs of real images and ground truth silhouettes to train a full encoder-decoder network, while keeping the decoder fixed. During this phase the encoder implicitly learns to predict SMPL [7] 3D body shape and pose parameters at the information bottleneck, since it is guided to correctly label the provided silhouette by the decoder. Note, as the second phase does not require a corresponding 3D body shape and pose for real images, several orders of magnitude larger datasets of human images and silhouettes can be used.

artificially generated 3D human body images (using SMPL [7]) and on real images from the Unite the People dataset [9].

The remainder of this paper is organised as follows. First, we provide an overview of related work in Section 2. We then describe our deep architecture used in experiments in Section 3. Finally, we proceed with experimental setup details and analysis in Sections 4 and 5 respectively.

2 Related Work

The work presented in this paper is closely related to three areas of computer vision and machine learning: state-of-the-art 3D human shape and pose prediction, structured deep learning and inverse graphics.

Human Shape and Pose Prediction. Similar to many areas of computer vision, state-of-the-art methods of human shape and pose prediction have benefitted substantially from recent adoption of Convolutional Neural Network (CNN) architectures. Advances have also been realised by employing complex neural network architectures [5, 13], additional predicted output modalities (e.g. part segmentation [14]), additional sensor modalities (e.g. Kinect [15]), statistical human body shape and pose modelling packages [7], and increasing availability of motion capture datasets (e.g. Human 3.6 [6], HumanEva [16]) as well as static image datasets [9]. However, as the complexity of predictions increase, obtaining sufficient data becomes a significant problem. Unlike current state-of-the-art methods in human body shape and pose prediction, our work allows for significant reduction of data labelling costs by incorporating indirect learning from graphics engines.

Structured Deep Learning. The advent of neural networks also gave a boost to complex deep structured models. Especially promising results were demonstrated by auto-encoders

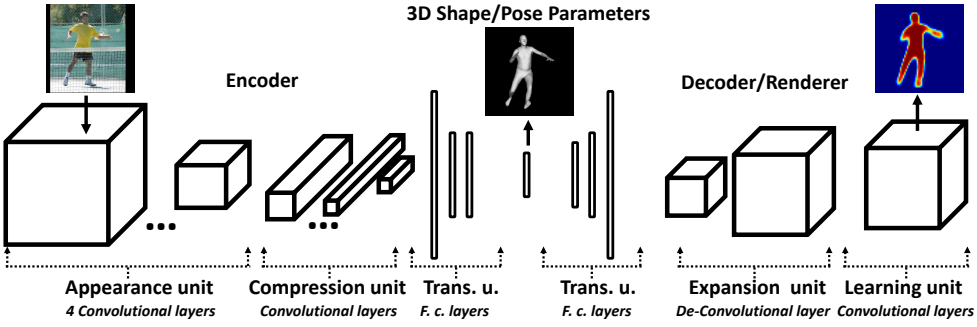

Figure 2: This figure illustrates key components of our proposed encoder-decoder network. For convenience of explanation, we split the encoder into *appearance*, *compression*, and *translation* units. The *appearance* unit learns convolutional filters for human silhouette and background separation. The *compression* unit further compresses the output of the appearance unit to a vector of dimensions $1 \times 1 \times 64$. The *translation* unit then converts this vector into shape and pose parameters using three fully connected layers. Similarly, the decoder is split into *translation*, *expansion* and *learning* units. The *translation* unit converts 3D shape and pose parameters into a low dimensional (9×9), 8 channel image via three fully connected layers and one reshape layer. The *expansion* unit uses a deconvolution layer to expand the output in both image dimensions (69×69) and number of channels (384). Finally, the *learning* unit uses a convolutional layer to compress this output to 2 channels.

(variational [7] or not) for face rendering [8], human pose prediction [13] and general multi-class segmentation [4]. Our proposed network architecture uses components of both structured prediction and generative modelling, but differs from most of the other works in this area due to the strict interpretability (i.e. each parameter has a pre-assigned meaning) of its prediction at the information bottleneck.

Inverse Graphics. Inverse graphics [1, 19] is the problem of recovering the underlying scene components by learning to render a target scene. Most of the approaches employ explicitly designed [11, 12] differentiable renderers with some exceptions, such as [8]. Our proposed decoder/renderer is fully learnt, differentiable and is simplified to predict pixel-wise 2D human body silhouette class labels as opposed to RGB values.

3 Model

In this section we provide a detailed description of our chosen encoder-decoder network architecture. While we use standard layers (e.g. convolution, fully connected, deconvolution) of CNNs, our grouping of them is non-standard. In particular, we split the encoder and decoder into three units each, serving particular purposes as described in the sections below and Figure 2.

3.1 Decoder

The decoder takes 82 SMPL [2] body shape and pose parameters (3 parameters for the whole body rotation, 10 shape parameters and 69 pose parameters) and predicts a two-class label image which corresponds to a silhouette generated by the input SMPL parameters. The proposed decoder is divided into three conceptual units: translation, expansion and learning.

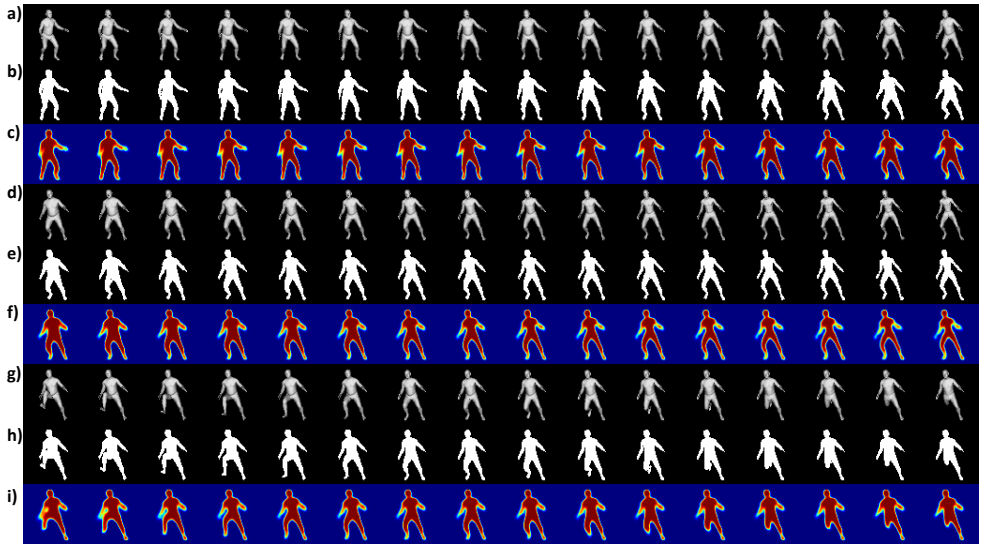

Figure 3: This figure shows the effect of varying one rotation (rows a-c), one shape (rows d-f) and one pose (rows g-i) parameter in the input to a trained decoder. Note that the first row of each triplet of consecutive rows corresponds to the rendered model, the second to the ground truth silhouette, and the third to the predicted silhouette. The decoder matches the variation of the input parameter well. Zoom in for the best view. See supplementary material for more results.

Translation unit. The translation unit converts 3D shape and pose parameters into a vector of length 648 to be reshaped into a grid by the subsequent expansion unit. This unit consists of 3 fully connected layers with 256, 384, and 648 outputs respectively.

Expansion unit. The expansion unit enlarges this image grid beyond the output silhouette image resolution and further increases the number of channels. This unit consists of a reshape unit outputting an 8 channel image of resolution 9×9 and a deconvolution layer further expanding that into a 69×69 image of 384 channels.

Learning unit. The learning unit compresses the expanded parameter space into the output format (a two-class label image of 64×64 pixel resolution). This unit consists of a convolutional filter (2 channels, kernel size 6) and a per-pixel softmax layer. Note that the relatively small output resolution is chosen to simplify the learning procedure of the decoder.

See Figure 3 for an illustration of the results of a trained decoder and its performance when the shape, pose and main rotation inputs are varied. Conceptually, the decoder behaves as a silhouette renderer. Also see Section 5.1 for a quantitative comparison of the performance of different variations of the proposed decoder architecture.

3.2 Encoder

The encoder takes an RGB image as an input and outputs 82 SMPL [2] body shape and pose parameters. As with the decoder, it is split into 3 conceptual units.

Appearance unit. The appearance unit learns appearance representations as in many common segmentation architectures. Here we use 7 pairs of convolutional and reLU layers. 4 max pooling layers are also used - these are situated after the 1st, 2nd, 4th and 6th

pairs of convolutional and reLU layers. The appearance unit takes an input of dimensions $256 \times 256 \times 3$ and produces an output of dimensions $16 \times 16 \times 256$. Note that while more complex choices for appearance units may improve performance, searching for them is beyond the scope of this work.

Compression unit. This unit further compresses the information coming from the appearance unit. Four sets of convolutional, reLU and max pooling layers reduce the output until it takes a shape of a 1×1 image of 64 channels. Note that such extreme compression is uncommon in segmentation networks due the risk of losing low level detail required for segmentation tasks. In this application, however, we aim to learn a highly compact vector representation as the output of the encoder, thus justifying such a design.

Translation unit. This unit converts the low dimensional convolutional filter output of the compression unit into the 3D body shape and pose model parameters. It uses 3 fully connected layers of 4096, 128, and 82 outputs respectively.

See Section 5.1 for a comparison of different variations of the proposed encoder architecture. Also see Section 5 and Figures 4, 5 for an explanation of the training procedure and test examples of the full network.

4 Data Preparation

We evaluated the proposed method using two datasets: an internal artificial dataset created from the SMPL [2] statistical body shape model and a subset of the Unite the People dataset [9]. Our data preparation process is as described below.

Artificial Images. To create the artificial data for both the encoder and decoder training we used the following procedure: first we selected[1] 10 shape, 20 pose and 3 main rotation parameters (limited to span a range of 72°) from the total parameter set of 82 parameters. We then sampled the selected shape and pose parameters from within the range of values observed in the Unite The People dataset [9]. For each data point, we created the corresponding RGB (used for full network training only) and silhouette image (used for both decoder and full network training). Two types of RGB images were created - one with low random pixel noise and one with high random pixel noise as well as artificial occlusions (modelled as occluding rectangles). When rendering, we positioned the 3D model at the centre of a square image with a 10% boundary. Examples of artificial images created can be found in row (a) of Figure 4.

Real Images. For our experiments with real images we used a subset of the Unite the People dataset [9] consisting of 8515 images with both 3D shape and pose parameters labels as well as ground truth segmentation. These images depict people in a wide variety of highly challenging poses. As in the previous section, we filtered out those images with main rotation parameters outside 72° range. This resulted in train and test datasets of 1307 and 139 examples respectively. We also cropped each RGB image and corresponding ground truth silhouette to leave a minimum of 10% background only boundary around the silhouette.

[1] Pose parameters were limited in order to decrease the networks sensitivity to the ill-posed nature of the problem as different parameter configurations may yield similar renderings of 3D models.

Architecture	Transl. Unit	Exp. Unit	Learn. Unit	Pred. silh. GA	CA	IoU
Simplified translation network	1 fc l.	deconv-384ch	1 conv. l.	0.91	0.82	0.73
Fully connected expansion network	3 fc l.	fc-69x69x96	1 conv. l.	0.94	0.89	0.81
Simplified learning net (deconv)	3 fc l.	deconv-2ch	NO	0.93	0.83	0.76
Simplified learning net (summation)	3 fc l.	deconv-384ch	2 ch. sum l.	0.95	0.89	0.83
1 Layer learning network	3 fc l.	deconv-384ch	1 conv. l.	0.95	0.90	0.84
2 Layer learning network	3 fc l.	deconv-384ch	2 conv. l.	0.95	0.91	0.85

(a)- Decoder

Architecture	App. Unit	Compr. Unit	Transl. Unit	Mean Lm. Dist Mean	Var	Pred. mod. silh. GA	CA	IoU
Simplified translation net	proposed	conv-1x1x64	1 fc l.	1.85	0.63	0.97	0.92	0.87
Original network	proposed	conv-1x1x64	3 fc l.	1.83	0.39	0.96	0.91	0.86
Deeper translation net	proposed	conv-1x1x64	4 fc l.	1.44	0.31	0.97	0.91	0.87
No compression network	proposed	NO	3 fc l.	2.37	0.37	0.95	0.88	0.82
Higher compression net	proposed	conv-1x1x16	3 fc l.	2.58	0.57	0.95	0.89	0.83

(b) - Encoder

Table 1: This table provides the quantitative evaluation of various decoder and encoder architectures. The architectures with bolded names are the ones used in the experiments reported in Section 5.3. The decoders are compared in terms of their predicted silhouette segmentation quality using the global accuracy (GA), class average (CA) accuracy and intersection over union (IoU) measures. The encoders are compared in terms of the quality of fit of 91 landmark points across the body [2] (mean and variance of average 3D distance per point is reported in native SMPL mesh distance unit scaled by a factor of 10) as well as the segmentation accuracy of the silhouette of the predicted model. The meaning of abbreviations used in the tables are as follows: *"1 fc l."* - one fully connected layer; *"deconv-384ch"* - deconvolutional layer outputting 384 channels; *"fc-69x69x96"* - fully connected layer with a total of 457056 outputs, reshaped to give a 69x69 image of 96 channels; *1 conv. l.* - one convolutional layer; *"2 ch. sum l."* - two channel sum layer which squeezes an N channel input into a two channel output by performing summation over half of the input channels each; *"conv-1x1x64"* a set of convolution layers which outputs a 1 pixel (1×1) image of 64 channels.

5 Experiment Results

In this section we describe the experimental setup and results of two groups of experiments. The first group of experiments compares different architecture choices for both the encoder and decoder on artificial data. The second group of experiments compares direct learning (3D shape and pose parameters provided at train time) and indirect learning on both artificial and real images.

5.1 Architecture Comparison

First we compare six different decoder architectures by making changes in one of the three units. Note that comparisons are chosen to highlight the design intuitions only, and one could almost certainly improve overall performance with a more complicated design. The details of the aforementioned six networks are provided in Table 1 (a) and are explained in more detail below.

- *Simplified translation network* - network obtained by reducing the number of fully connected layers from 3 to 1 (with 648 outputs) in the translation unit.
- *Fully connected expansion network* - network obtained by replacing the deconvolutional layer in the expansion unit with a fully connected layer (457056 outputs). Note that due to limited GPU memory, the number of channels for this new expansion unit was reduced to 96.
- *Simple learning network with deconvolution* - network obtained by replacing the learning unit with a crop layer and reducing the number of channels of the deconvolutional layer in the expansion unit to 2, so as to match the decoder output format.
- *Simple learning network with summation* - network obtained by replacing the convolutional layer in the learning unit with a summation layer which outputs a 64×64 image of two channels, each of which is formed by a summation over half (194) of input channels.
- *1 Layer learning* - the original model described in Section 3.1 which contains 1 convolutional layer in the learning unit.
- *2 Layer learning* - a network obtained by adding an additional convolutional layer (128 channels, kernel of size 5, stride 1) to the start of the learning unit.

All the aforementioned models were trained for 500K iterations (batch size 1) on an artificial image dataset of 100K training samples. Pairs of human body parameters and silhouettes were obtained using the procedure described in Section 4. An additional 1000 artificial images were used to measure the segmentation quality in terms of global accuracy, class average accuracy and average intersection over union. As shown in Table 1 (a) our original choice of "1 Layer learning" outperforms all but one architecture - "2 Layer learning". As the latter architecture is more complex but yields only modest benefit, we perform all further experiments with the decoder described in Section 4.

As with the decoder we perform a quantitative evaluation of 5 different encoder architectures. These were all trained using the same procedure (100K high noise training samples, 500K iterations, batch size 4) and have the same fixed decoder chosen via experimentation as described above. These architectures are described in Table 1 (b) and below:

- *Simplified translation network* - a network obtained by reducing the number of fully connected layers from 3 to 1 (82 outputs) in the translation unit.
- *Original network* - a network using the architecture proposed in Section 3.
- *Deeper translation network* - a network obtained by adding an additional inner product layer (256 outputs) into the translator unit.
- *No compression network* - a network obtained by removing the compression unit.
- *Higher compression network* - a network obtained by reducing the output size of the compression unit to 16 channels.

The aforementioned encoder architectures are compared by predicted model silhoutte segmentation accuracy (as opposed to predicted silhouette accuracy) and average 91 landmark point distance in 3D as shown in Table 1 (b). Note that all architectures have similar predicted silhouette accuracy. However, our proposed network is outperformed by one architecture - "Deeper translation net". This suggests a potential for further exploration to arrive at even more accurate model, yet is out of the scope of this work.

	Method	Exp Type	Encoder Noise	Occ	# of Samples	Mean Lm. Dist. Mean	Var	Pred. Mod. Silh. GA	CA	IoU	Pred. Silh. GA	CA	IoU
a)	Direct Learning	A	Low	N	100K	1.22	0.14	0.94	0.87	0.79	0.95	0.89	0.82
b)	Indirect Learning	A	Low	N	100K	1.39	0.29	0.97	0.92	0.88	0.97	0.95	0.90
c)	Direct Learning	A	High	Y	100K	1.23	0.13	0.94	0.87	0.79	0.95	0.89	0.82
d)	Indirect Learning	A	High	Y	100K	1.83	0.39	0.96	0.91	0.86	0.97	0.94	0.89
e)	Ground Truth Models [9]	R	N/A	N/A	N/A	N/A	N/A	0.95	0.88	0.82	N/A	N/A	N/A
f)	Direct Learning	R	Low	N	101.3K	1.23	0.46	0.92	0.82	0.73	0.92	0.84	0.74
g)	Mixed Learning	R	Low	N	101.3K	1.31	0.43	0.93	0.84	0.75	0.93	0.85	0.77
h)	Indirect Learning	R	Low	N	101.3K	1.89	0.18	0.94	0.87	0.79	0.94	0.88	0.80
i)	Direct Learning	R	High	Y	101.3K	1.05	0.36	0.91	0.80	0.71	0.92	0.82	0.73
j)	Mixed Learning	R	High	Y	101.3K	1.43	0.47	0.93	0.84	0.76	0.94	0.86	0.78
k)	Indirect Learning	R	High	Y	101.3K	1.90	0.22	0.95	0.89	0.83	0.96	0.92	0.86

Table 2: This table shows the full network training results for both artificial (rows a-d) and real (rows f-k) datasets. Network training on triplets (loss applied at both body shape and pose parameters and silhouette prediction layers) of image data, body shape and pose model parameters and silhouette, referred to as "direct learning" (rows a, c, f, i) is compared with our proposed "indirect learning" approach (rows b, d, h, k). "Mixed learning" (rows g, j) corresponds to a training procedure in which 20% of the training data is labelled with body pose and shape parameters. The networks are compared in their quality of fit for 91 3D model landmarks as well as silhouette fit of both the directly predicted silhouette and as well as the silhouette corresponding to the predicted model parameters. Silhouette fit is evaluated in global pixel accuracy, class average accuracy, and average intersection over union for foreground and background classes.

5.2 Artificial Data

In this section we compare our proposed model against the more standard *direct* learning approach which uses body shape and pose parameters at train time. For this purpose we use two artificial datasets (low noise and high noise) of 100K images each created using the procedure described in Section 4.

Both indirect learning and direct learning approaches use the same architecture proposed in Section 3, the only difference being that an additional Euclidean loss is applied on the layer which predicts the shape and pose parameters when training in a direct learning scenario. The same decoder trained on 100K images for 1.5M iterations is loaded into each network and then fixed. Each encoder is then trained for 500K iterations.

Rows a, b and c, d in Table 2 report results when trained on low noise and high noise datasets respectively for both networks. As in Table 1 models were evaluated using 3D model fit and segmentation (both predicted silhouette and predicted model silhouette) quality metrics. As shown in the table, both the segmentation and the 3D model prediction slightly deteriorates when the network is trained on the highly noisy images. Also as expected, the indirectly trained network achieves a higher quality segmentation, whereas the direct trained network achieves a higher 3D model fit score. Nevertheless, it can be seen qualitatively from Figure 4 that the indirect method is still able to achieve highly accurate model fits.

5.3 Real Image Data

For experiments with images from the Unite the People dataset [9], we performed a similar comparison of direct and indirect learning (Table 2, rows f and h). Note that in this case we created datasets containing both artificial (100K) and real images (1.3K). For decoder training, we reused the same decoders as in the previous section. For full network training, we split the training procedure into two stages: we first trained on (i) a reduced dataset of 11.3K images (1.3K real) for 400K iterations and then on (ii) the full 101.3K dataset for a

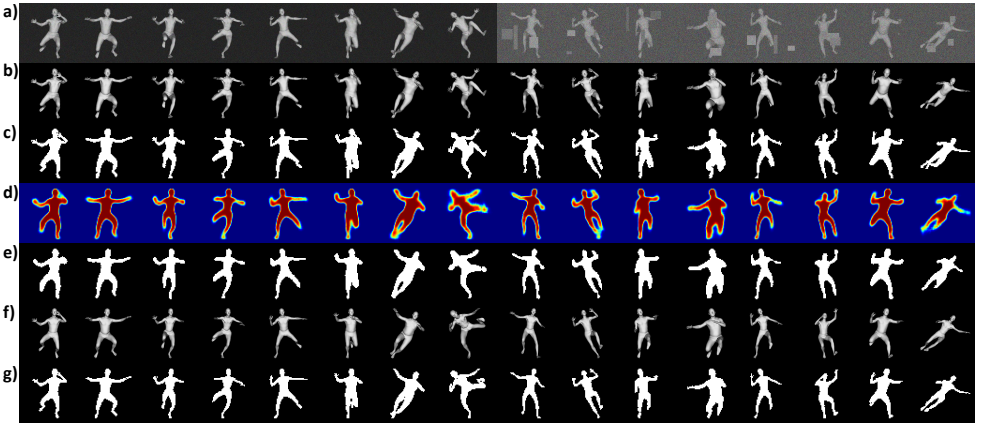

Figure 4: This figure illustrates test results on artificial data created using SMPL [2] model using low random noise (columns 1-8) and high random noise with artificial occlusion (columns 9-16) from networks reported in rows (b) and (d) of Table 2 respectively. Rows (a-c) show input images, corresponding source models and generated silhouettes.Rows (d-e) show predicted silhouette per pixel probabilities and corresponding silhouettes obtained by choosing the most likely foreground/background prediction per pixel. Finally rows (f) and (g) show rendered images of predicted human models and corresponding silhouettes. Note how well the predicted models and silhouettes match the ground truth images even when occlusions are present. Zoom in for the best view. See supplementary material for more results.

final 100K iterations. For both stages, a batch size of 4 was used. This split was introduced in order to avoid an extreme over-fit to the artificial examples. Also note that we further evaluated a "mixed learning" scenario (Table 2, rows g and j) where 20% of both the artificial and real images had corresponding 3D model parameters provided.

As shown in Table 2, similar to the artificial data experiments, networks trained with a higher degree of indirectness exhibited higher segmentation performance but lower 3D model fitting quality. On the other hand, as shown in Figure 5, a relatively good fit on real images is obtained despite the complete lack of real image and corresponding shape and pose parameter pairs in the training process.

6 Summary

In this work we presented a novel method for the indirect training of deep networks for structured prediction of 3D human shape and pose. Unlike most state-of-the-art approaches, our method does not require hard-to-obtain 3D human shape and pose labels for training on real world images, but instead leverages the power of a decoder trained on artificial data. It demonstrates high accuracy on artificial images and while its accuracy deteriorates on real-world images, close fits to ground truth can still be recovered even when using training procedures with no exposure to any real image and corresponding shape and pose parameter pairs. We hope that more complex network architectures and additional high quality training data will enable our proposed method to compete with state-of-the-art direct learning approaches.

Figure 5: This figure illustrates test results on a portion of the Unite the People dataset [9]. Both train and test images were filtered as explained in Section 4. Rows (a-d) show input images, hand labelled silhouettes, corresponding source models and generated silhouettes. Rows (e-f) show predicted silhouette per pixel probabilities and corresponding silhouettes obtained by choosing the most likely foreground/background prediction per pixel. Finally rows (f) and (g) show rendered images of predicted human models and corresponding silhouettes. Despite the lack of pose and shape parameter data in the training of encoder, in the majority of the cases the overall posture is correct with some errors particularly on the configuration of the hands and legs. Zoom in for the best view. See supplementary material for more results.

References

[1] B. G. Baumgart. *Geometric Modeling for Computer Vision.* PhD thesis, Stanford, USA, 1974.

[2] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision (ECCV)*, October 2016.

[3] R. Girshick. Fast R-CNN. In *Proceedings of the International Conference on Computer Vision (ICCV)*, December 2015.

[4] S. Hong, H. Noh, and B. Han. Decoupled deep neural network for semi-supervised semantic segmentation. In *Advances in Neural Information Processing Systems 28*. 2015.

[5] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision (ECCV)*, October 2016.

[6] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, Jul 2014.

[7] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

[8] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum. Deep convolutional inverse graphics network. In *Advances in Neural Information Processing Systems 28*. 2015.

[9] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[10] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, September 2014.

[11] M. M. Loper and M. J. Black. OpenDR: An approximate differentiable renderer. In *European Conference on Computer Vision (ECCV)*, September 2014.

[12] V. K. Mansinghka, T. D. Kulkarni, Y. N. Perov, and J. B. Tenenbaum. Approximate bayesian image interpretation using generative probabilistic graphics programs. In *Annual Conference on Neural Information Processing Systems (NIPS)*, December 2013.

[13] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision (ECCV)*, October 2016.

[14] G. Oliveira, A. Valada, C. Bollen, W. Burgard, and T. Brox. Deep learning for human part discovery in images. In *IEEE International Conference on Robotics and Automation (ICRA)*, May 2016.

[15] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, and A. Blake. Efficient human pose estimation from single depth images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2821–2840, Dec 2013.

[16] L. Sigal, A. O. Balan, and M. J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87(1):4, 2009.

[17] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[18] B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, and P. Fua. Structured prediction of 3d human pose with deep neural networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, September 2016.

[19] D. Terzopoulos, J. Pltt, A. Barr, D. Zeltzer, A. Witkin, and J. Blinn. Physically-based modeling: Past, present, and future. *SIGGRAPH Comput. Graph.*, 23(5), 1989.