Deep Autoencoder for Combined Human Pose Estimation and Body Model Upscaling

Matthew Trumble[1], Andrew Gilbert[1], Adrian Hilton[1], John Collomosse[1,2]

[1]Centre for Vision Speech and Signal Processing,
University of Surrey
[2]Creative Intelligence Lab,
Adobe Research

Abstract. We present a method for simultaneously estimating 3D human pose and body shape from a sparse set of wide-baseline camera views. We train a symmetric convolutional autoencoder with a dual loss that enforces learning of a latent representation that encodes skeletal joint positions, and at the same time learns a deep representation of volumetric body shape. We harness the latter to up-scale input volumetric data by a factor of 4×, whilst recovering a 3D estimate of joint positions with equal or greater accuracy than the state of the art. Inference runs in real-time (25 fps) and has the potential for passive human behaviour monitoring where there is a requirement for high fidelity estimation of human body shape and pose.

Keywords: Deep Learning, Pose Estimation, Multiple Viewpoint Video

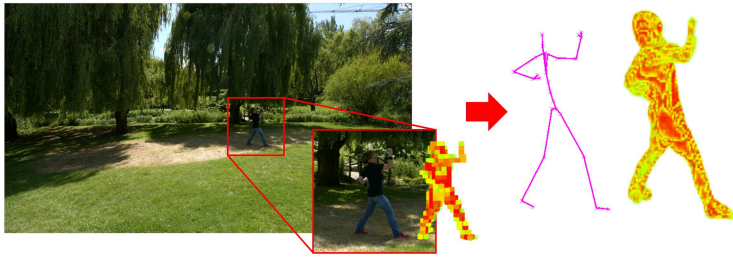

Fig. 1. Simultaneous estimation of 3D human pose and 4× upscaled volumetric body shape, from coarse visual hull data derived from a sparse set of wide-baseline views.

1 Introduction

Multiple viewpoint video of open spaces (e. g. for sports or surveillance) is often captured using a sparse set of wide-baseline static cameras, in which human subjects are relatively small (tens of pixels in height) due to their physical distance. Nevertheless, it is useful to infer human behavioural data from this limited knowledge for performance analytics or security. In this paper, we explore the possibility of using a deeply learned prior inferring high fidelity three-dimensional (3D) body shape and skeletal pose data from a coarse (low-resolution) volumetric estimate of body shape estimated across a sparse set of camera views (Fig. 1).

The technical contribution of this paper is to explore the possibility of learning a deep representation for volumetric (3D) human body shape driven by a latent encoding for the skeletal pose that can, in turn, be inferred from coarse volumetric shape data. Specifically, we investigate whether convolutional autoencoder architectures, commonly applied to 2D visual content for de-noising and up-scaling (super-resolution), may be adapted to up-scale volumetric 3D human shape whilst simultaneously providing high-level information on the 3D human pose from the bottle-neck (latent) representation of the autoencoder. We propose a symmetric autoencoder with 3D convolutional stages capable of refining a probabilistic visual hull (PVH) [1] i. e. voxel occupancy data derived at very coarse scale (grid resolution $32 \times 32 \times 32$ encompassing the subject). We demonstrate that our autoencoder is able to estimate an up-scaled body shape volume at up to $128 \times 128 \times 128$ resolution, whilst able to estimate the skeleton joint positions of the subject to equal or better accuracy than the current state of the art methods due to deep learning.

2 Related Work

Our work makes a dual contribution to two long-standing Computer Vision problems: super-resolution (SR) and human pose estimation (HPE).

Super-resolution: Data-driven approaches to image SR integrate pixel data e. g. from auxiliary images [2], or from a single image [3, 4]) to perform image up-scaling or restoration. Model based approaches learn appearance priors from training images, applying these as optimization constraints to solve for SR content [5]. A wide variety of machine learning approaches have been applied to the latter e. g. sparse coding [6], regression trees [7], and stacked autoencoders [8]; many such approaches are surveyed in [9]. Deep learning has more recently applied convolutional autoencoders for up-scaling of images [10–12] and video [13]; our work follows suit, extending symmetric autoencoders commonly used for image restoration to volumetric data using 3D (up-)convolutional layers [14]. Our work is not the first to propose volumetric super-resolution. Data-driven volumetric SR has been explored using multiple image fusion across the depth of field in [15] and across multiple spectral channels in [6]. Very recent work by Brock *et al.* explores deep variational auto-encoders for volumetric SR of objects [16]. However, our work is unique in its ability to upscale to $4\times$ whilst simultaneously estimating human pose to a high accuracy, exploiting a learned latent representation encoding joint positions.

Human pose estimation has been classically approached through top-down fitting of models such as Pictorial structures [17], fused with Ada-Boost shape classification in [18]. Conditional dependencies between parts (limbs) during model fitting were explored in [19, 20]. Huang [21] tracked 3D mesh deformation over time and attach a skeleton to tracked vertices. The *SMPL* body model [22] provides a rich statistical body model that can be fitted to (possibly incomplete) visual data. Marcard [23] explored the orthogonal modality of IMU measurements using SMPL for HPE without visual data. Malleson [24] used IMUs with a full

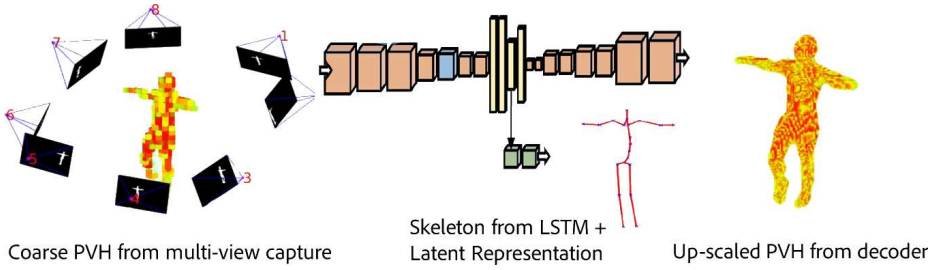

Fig. 2. Overview of the proposed method. A coarse PVH is estimated as input volumetric data (32^3 voxels) and up-scaled via tricubic interpolation to a $(32n)^3$ voxel grid (where $n = \{1, 2, 4\}$). The input PVH is deeply encoded to the latent feature representation (3D joint positions). Non-linear decoding of the feature via successive up-convolutional layers yields a higher fidelity PVH of $(32n)^3$ voxels.

kinematic solve to estimate 3D pose. SMPL was recently applied to a deep encoder-decoder network to estimate 3D pose from 2D images [25]. Several deep approaches estimate 2D pose or infer 3D pose from intermediate 2D estimations. DeepPose [26] applies a convolutional neural network (CNN) cascade. Descriptors learned via CNN have been used in 2D pose estimation from low-resolution 2D images [27] and real-time multi-subject 2D pose estimates were demonstrated by cao [28]. Sanzari [29] estimates the location of 2D joints, before predicting 3D pose using appearance and probable 3D pose of parts. Zhou [30] integrates 2D, 3D and temporal information to account for uncertainties in the data.

The challenge of estimating 3D human pose from volumetric data is more sparsely explored. Trumble [31] used a spherical histogram and later voxel input to regress a pose estimate using a CNN [32]. Pavlakos [33] used a simple volumetric representation in a 3D convnet for pose estimation. While Tekin [34] included a pretrained autoencoder within the network to enforce structural constraints. Our work also trains an autoencoder for HPE but simultaneously infers a high resolution body model via a dual loss function.

3 Estimating Human Pose and Body Shape

Our method accepts a coarse resolution volumetric reconstruction of a subject as input, and in a single inference, step estimates both the skeletal joint positions and a higher resolution (up-scaled) volumetric reconstruction of that subject (Fig. 2). Sec. 3.1 first describes how the input volumetric reconstruction is formed, through a simplified form of Graumann 's probabilistic visual hull (PVH) [1]. The architecture of our 3D convolutional autoencoder is then described in Sec. 3.2 including the dual loss function necessary to learn a deep representation of body shape and the latent pose representation. Finally, Sec. 3.3 describes the data augmentation and methodology for training the network.

3.1 Volumetric Representation

The capture volume $\mathcal{V} \in \mathbb{R}^3$ containing the subject is observed by a set of C calibrated cameras $c = [1, C]$ for which camera world position T_c and orientation

R_c (both matrices in homogeneous form) are known as are intrinsics: camera focal length (f_c) and optical center $[o_c^x, o_c^y]$. An external process (e. g. a person tracker) is assumed to isolate the bounding sub-volume $X_I \in \mathcal{V}$ corresponding to, and centered upon, a single subject of interest, and which is decimated to a coarse voxel grid $V = \{v_x^i, v_y^i, v_z^i\}$ for $i = [1, ..., 32^3]$ where V denotes the coarse voxel volume passed as input to the network in Sec 3.2. Each voxel $v^i \in V$ projects to coordinates $(x[v^i], y[v^i])$ in each camera view c derived in homogeneous form via pin-hole projection:

$$\begin{bmatrix} \alpha x[v^i] \\ \alpha y[v^i] \\ \alpha \end{bmatrix} = \begin{bmatrix} f_c & 0 & o_c^x & 0 \\ 0 & f_c & o_c^y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \left(-R_c^{-1} T_c\right) \begin{bmatrix} v_x^i \\ v_y^i \\ v_z^i \\ 1 \end{bmatrix}. \tag{1}$$

Given a soft matte I_c obtained, for example by background (clean-plate) subtraction, the probability of the voxel being part of the performer in a given view c is:

$$p(v^i|c) = I_c(x[v^i], y[v^i]). \tag{2}$$

The overall probability of occupancy for a given voxel $p(v^i)$ is:

$$p(v^i) = \prod_{i=1}^{C} 1/(1 + e^{p(v^i|c)}). \tag{3}$$

For all voxels $v^i \in V$ we compute $p(v^i)$ to form the coarse input PVH.

3.2 Dual Loss Convolutional Autoencoder

We use a convolutional autoencoder with a symmetrical 'hourglass' (encoder-decoder) architecture. The goal of the network is learn a deep representation given an input tensor $\mathbf{V_I} \in \mathbb{R}^{N \times N \times N \times 1}$ encoding the coarse PVH, V at a given resolution $N = (32n)^3$, where $n = \{1, 2, 4\}$ is a configuration parameter determining the degree of up-scaling required from the network $(1\times, 2\times, 4\times)$ respectively. The coarse PVH input V is scaled via tri-cubic interpolation to fit $\mathbf{V_I}$. We train the deep representation to solve the prediction problem $\mathbf{V_H} = \mathcal{F}(\mathbf{V_I})$ for similarly encoded output tensor $\mathbf{V_O}$, where

$$\mathbf{V_O} = \mathcal{F}(\mathbf{V_I}) = \mathcal{D}(\mathcal{E}(\mathbf{V_I})) \tag{4}$$

for the end to end trained encoder ($\mathcal{E}$) and decoder ($\mathcal{D}$) functions The encoder yields a latent feature representation via a series of 3D convolutions, max-pooling and fully-connected layers. We enforce $J(V_I) = \mathcal{E}(\mathbf{V_I})$ where $J(\mathbf{V_I})$ is a skeletal pose vector corresponding the input PVH; specifically a 78-D vector concatenation of $26\times$ 3D Cartesian joint coordinates in $\{x, y, z\}$. The decoder half of the network inverts this process to output tensor $\mathbf{V_O}$ matching the input resolution but with higher fidelity content. Fig. 3 illustrates our architecture which

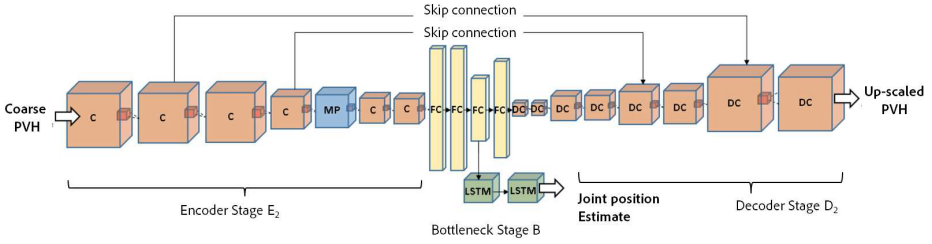

Fig. 3. Proposed convolutional autoencoder structure. The coarse input PVH is encoded into a latent feature representation via 3D (C)onvolutional, (M)ax-(P)ooling and (F)ully-(C)onnected layers. The decoder uses the latent representation to synthesize an up-scaled PVH via (D)e-(C)onvolutional layers. Two skip connections bridge the latent representation which is constrained during training to encode Cartesian joint positions. During inference these are passed through an LSTM to enhance temporal consistency to produce the joint position skeleton estimate. Architecture pictured here is for 2× scale-up – in order to accommodate different receptive field sizes for V_I/V_O (de-)convolutional layer count is adjusted – see Tbl. 1.

incorporates two skip connections bypassing the network bottleneck to allow the output from a convolutional layer in the encoder to feed into the corresponding up-convolution layer in the decoder. Activations from the preceding layer in the main network and skip connection data are combined via mean average rather than channel augmentation/residuals.

Tbl. 1 describes the parameters (filter count and size) of each layer. We report experiments up-scaling to $n = \{1, 2, 4\}$ requiring varying sizes of receptive field to accommodate $\mathbf{V_I}$ and $\mathbf{V_O}$. For each step up in scale, we add a single additional convolutional layer to the encoder, and two additional de-convolutional layers to the decoder. Max-pooling occurs always at the fourth convolutional layer, and the filter size is $3 \times 3 \times 3$ except for the first two and last two layers, where the filter size is $5 \times 5 \times 5$.

Learning the end-to-end mapping from coarse PVH to both an up-scaled PVH and accurate 3D joint positions requires estimation of the weights ϕ in $\mathcal{F}$ represented by the convolutional and deconvolutional kernels.

Specifically, given a collection of M training triplets $\{\hat{\mathbf{V}}_\mathbf{I}, \hat{\mathbf{V}}_\mathbf{O}, \hat{J}\}$, where $p^i \in \hat{\mathbf{V}}_\mathbf{I}$ is voxel data from a coarse (input) PVH, $q^i \in \hat{\mathbf{V}}_\mathbf{O}$ is voxel data of an ideal up-scaled PVH, and j is a vector of ideal joint positions for the given volume. We minimize the Mean Squared Error (MSE) at the outputs of the bottleneck and decoder stages across $M = N \times N \times N$ voxels:

$$\mathcal{L}(\phi) = \frac{1}{M} \sum_{i=1}^{M} \|\mathcal{F}(p^i : \phi) - q^i\|_2^2 + \lambda \|\mathcal{E}(\hat{\mathbf{V}}_\mathbf{I} : \phi) - j\|_2^2. \tag{5}$$

These training triplets are formed by extracting voxel volumes from exemplar multi-view video footage at resolution $N \times N \times N$ (yielding $\hat{\mathbf{V}}_\mathbf{O}$ and the artificially down-sampling to $32 \times 32 \times 32$ to yield V (from which $\mathbf{V_I}$ is up-sampled via tri-cubic interpolation). Human pose (joint positions) corresponding to the

Network Stage	#Layers	#Channels/Layer
E_1	5	96* 96* 96 96-M 96
E_2	6	32* 64* 96 96-M 96 96
E_4	7	32* 32* 32 64-M 96 96 96
B	4	1024 1024 78-J 216
D_1	6	96 96 96 96 64* 1*
D_2	8	96 96 96 96 64 64 32* 1*
D_4	10	96 96 96 96 64 64 32 32 32* 1*

Table 1. Convolution layer parameters for the encoder (E_n), bottleneck (B), and decoder (D_n) stages for $n = \{1, 2, 4)\times$. Suffix $-M$ indicates max-pooling. All E_n and D_n layers learn $3 \times 3 \times 3$ filters, except where indicated by $*$ filters are $5 \times 5 \times 5$. All B layers are fully-connected including the latent representation (3D joint positions) suffixed $-J$.

multi-view video frame is acquired using a commercial (Vicon Blade) human performance capture system run in parallel with video acquisition (such annotations are provided with the *TotalCapture* and *Human3.6M* datasets).

3.3 Training Methodology

To train $\mathcal{F}$ we use Adadelta [35] an extension of Adagrad, with the pose term of the dual loss (eq. 5) scaled by a factor of λ. We found the approach insensitive to this parameter up to an order of magnitude setting $\lambda = 10^{-3}$ for all experiments. Below 10^{-3}, the bottleneck convergences to a semantic representation of the pose that is stable but does not resemble joint angles — above 10^{-2} the network will not converge. Data is augmented during training with a random rotation around the central vertical axis of the PVH. Before full network training, the encoder stage is trained separately, purely as a pose regression task, up to the latent representation layer. These trained weights initialize the encoder stage to help constrain the latent representation during full, dual-loss network training. Training typically converges within 100 epochs.

3.4 Enforcing Temporal Consistency

Given the rich temporal nature of the pose sequences, it is prudent to exploit and enforce the temporal consistency of the otherwise detection based human joint estimation. By enforcing temporal consistency it is possible to smooth noise in individual joint detections that otherwise would cause large estimation errors. To learn a model of the motion over time we employ Long Short Term Memory (LSTM) layers [36], they have been heavily utilized in applications where long term temporal correlation can be exploited such as *e.g.* speech recognition [37], video description [38], and pose estimation [39]. LSTM layers are based on a recurrent neural network (RNN). They can store and access information over long periods of time but are able to mitigate the vanishing gradient problem common in RNNs through a specialized gating mechanism. The input vector from the encoder $\mathbf{J}_i(t) = \mathcal{E}(\mathbf{V_I})$ at time t consisting of concatenated joint spatial coordinates is passed through a series of gates resulting in an output joint vector $\mathbf{J}_o(t)$. The aim is to learn the function that minimizes the loss between the

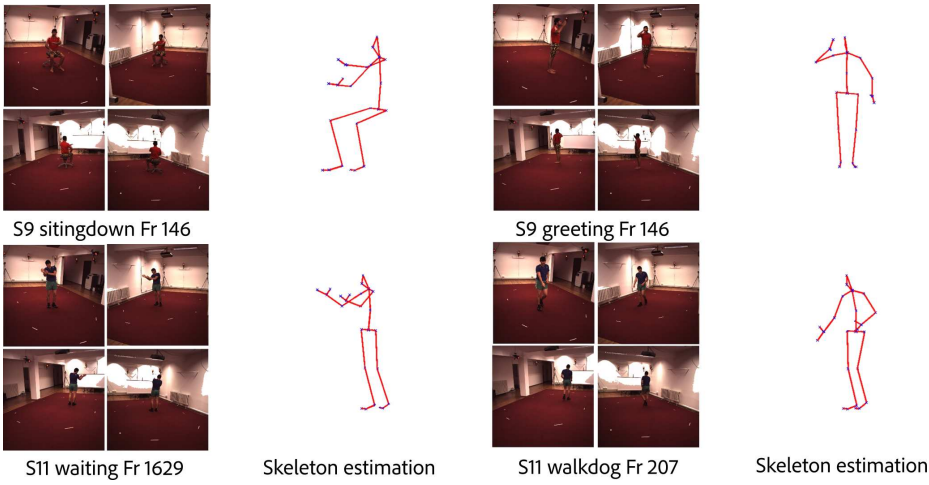

Fig. 4. Representative visual results for pose estimation on Human 3.6M across four test sequences (source footage from four views and inferred 3D skeletal pose).

input vector and the output vector $\mathbf{J}_o = o_t \circ tanh(c_t)$ ($\circ$ denotes the Hadamard product) where o_t is the output gate, and c_t is the memory cell, a combination of the previous memory c_{t-1} multiplied by a decay based forget gate, and the input gate. Thus, intuitively the LSTM result is the combination of the previous memory and the new input vector. In the implementation, our model consists of two LSTM layers both with 1024 memory cells, using a look back of $f = 5$.

4 Evaluation and Discussion

To quantify the improvement in both the upscaling of low resolution volumetric representations and human pose estimation, we evaluate over three public multi-view video datasets of human actions. For *Human 3.6M* [40] we estimate the 3D human pose, and examine the performance of the skeleton estimation and volume upscaling in the *TotalCapture* [32] dataset. Finally, we visualize the results of the skeleton estimation and upscaling on the dataset *TotalCaptureOutdoor* [41], a challenging collection of multi-view human actions shot outdoors.

4.1 Human 3.6M evaluation

The 3D human pose estimation dataset *Human3.6M* [40] is a 4 camera view dataset of 10 subjects performing 210 actions at 50Hz in a 360° arrangement. A 3D ground truth for joint positions (key points) are available via annotation using a commercial marker-based motion capture system, allowing quantification of error. The dataset consists of 3.6 million video frames, balanced over 5 female and 6 male subjects. They perform common activities such as posing, sitting and

Approach	Direct.	Discus	Eat	Greet.	Phone	Photo	Pose	Purch.
Lin [42]	132.7	183.6	132.4	164.4	162.1	205.9	150.6	171.3
ekin [43]	85.0	108.8	84.4	98.9	119.4	95.7	98.5	93.8
Tome [44]	65.0	73.5	76.8	86.4	86.3	110.7	68.9	74.8
Trumble [32]	92.7	85.9	72.3	93.2	86.2	101.2	75.1	78.0
Lin [45]	58.0	68.3	63.3	65.8	75.3	93.1	61.2	65.7
Martinez [46]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1
Proposed	41.7	43.2	52.9	70.0	64.9	83.0	57.3	63.5
	Sit.	Sit D	Smke	Wait	W.Dog	walk	W. toget.	Mean
Lin [42]	151.6	243.0	162.1	170.7	177.1	96.6	127.9	162.1
ekin [43]	73.8	170.4	85.1	116.9	113.7	62.1	94.8	100.1
Tome [44]	110.2	173.9	85.0	85.8	86.3	71.4	73.1	88.4
Trumble [32]	83.5	94.8	85.8	82.0	114.6	94.9	79.7	87.3
Lin [45]	98.7	127.7	70.4	68.2	73.0	50.6	57.7	73.1
Martinez [46]	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
Proposed	61.0	95.0	70.0	62.3	66.2	53.7	52.4	62.5

Table 2. A Comparison of our approach to other works on the Human 3.6m dataset

giving directions. To allow comparison to other approaches we follow the same data partition protocol as in previous works [40, 42, 43, 45, 44, 46], and we use the publicly released foreground mattes. The training data consists of subjects S1, S5, S6, S7, S8 and it is tested on unseen subjects S9, S11. We compare our approach to many previously published state of the art methods, using 3D Euclidean (L_2) error to compute accuracy. The error is measured between each ground truth and estimated 3D joint position and is averaged over all frames and all 17 joints in millimetres (mm). The results of our approach are evaluated qualitatively in Fig 4 and quantitatively in Tbl. 2, drawing a comparison to state of the art approaches.

Our approach outperforms with the lowest mean joint error on the challenging Human3.6M dataset, slightly reduced over the state of the art approach by Martinez [46], with a similar mean joint error of just over 6cm. This is averaged over both test subjects and the 59 sequences. The error decrease over the other approaches is possible due to the dual loss formulation ensuring that the skeleton is kept bounded by realistic 3D volume representations after the decoder. Our approach struggles with the actions Sit Down and Photo, the action sit down contains a chair and given the already poor quality of the PVH it is likely that such incorrect joint estimations occur. In the sequences of *photo* the hands of the subject are close the subject head and it is likely the PVH volume doesn't contain enough discriminative information to correctly estimate their location. However, despite these two sequences, all others have a low error score and are smooth and qualitatively realistic. We show qualitative comparisons with respect to the ground truth in Fig. 4. To illustrate the stability of our approach across different test subjects we performed five rounds of cross-validation using multiple pairs of test subjects with the remaining subjects held out for training the model. Table 3 shows the standard test on S9 and S11 (mean accuracy of 62.5mm) from Table 2 against the mean and standard deviation from our cross-validation experiment. The mean performance across random pairs of test subjects is similar to that

Approach	Direct.	Discus	Eat	Greet.	Phone	Photo	Pose	Purch.
CrossVal mean	52.2	49.8	53.0	63.1	61.4	76.8	63.2	59.3
CrossVal sd	7.6	5.1	9.1	5.8	3.9	4.7	10.4	6.9
Proposed	41.7	43.2	52.9	70.0	64.9	83.0	57.3	63.5
	Sit.	Sit D	Smke	Wait	W.Dog	walk	W. toget.	Mean
CrossVal mean	64.9	108.3	68.9	63.0	63.6	57.4	55.0	70.2
CrossVal sd	5.2	15.8	5.7	3.2	6.9	5.2	3.0	3.3
Proposed	61.0	95.0	70.0	62.3	66.2	53.7	52.4	62.5

Table 3. A Comparison of testing on subjects S9 and S11 against a five-fold cross validation of other subject pairs on the Human 3.6m dataset

of the official S9/S11 test split, and the σ is low. Thus they serve to show the stability of the approach across different test subject pairings.

4.2 TotalCapture evaluation

Approach	SeenSubjects(S1,2,3)			UnseenSubjects(S4,5)			Mean
	W2	FS3	A3	W2	FS3	A3	
Tri-CPM-LSTM [28]	45.7	102.8	71.9	57.8	142.9	59.6	80.1
2D Matte-LSTM [31]	94.1	128.9	105.3	109.1	168.5	120.6	121.1
Trumble [32]	30.0	90.6	49.0	36.0	112.1	109.2	70.0
AutoEnc-Front-Half	42.0	120.5	59.8	58.4	162.1	103.4	85.4
AutoEnc-x1-LSTM	15.1	54.8	26.6	25.9	76.0	42.7	38.6
AutoEnc-x2-LSTM	13.0	47.0	23.0	21.8	68.5	40.9	34.1
AutoEnc-x4-LSTM	13.4	49.8	24.3	22.0	71.7	40.7	35.5

Table 4. Comparison of our approach on TotalCapture to other human pose estimation approaches, expressed as average per joint error (mm).

In addition, we evaluate our approach on the TotalCapture dataset [32]. This is also a 3D human pose estimation dataset with the ground truth joint position provided by Vicon markers. It is also captured indoors in a volume roughly measuring 8x4m with 8 calibration HD video cameras at 60Hz in a 360°. There are a total of 5 subjects performing 4 actions with 3 repetitions at 60Hz in a 360° arrangement. There are publicly released foreground mattes that we use as the input to our approach. Note to provide the Vicon groundtruth the subjects in both TotalCapture and Human3.6M are wearing dots visible to infrared cameras. However these dots are not used explicitly by our algorithm, and their size is negligible compared to the performance volume. There are five subjects in the dataset, four male, and one female, each performs four diverse performances, that are repeated 3 times: *ROM, Walking, Acting, and Freestyle*. The length of each sequence is between 3000-5000 frames, this results in a total of $\sim 1.9M$ frames of synchronized groundtruth video data. Especially within the acting and freestyle sequences, there is great diversity in the actions performed, as illustrated in the qualitative results in Fig. 5. To allow for comparison between seen and unseen subjects in the test evaluation, the test consists of sequences Freestyle3 (**FS3**), Acting (**A3**) and Walking2 (**W2**) on subjects 1,2,3,4 and 5. While the training

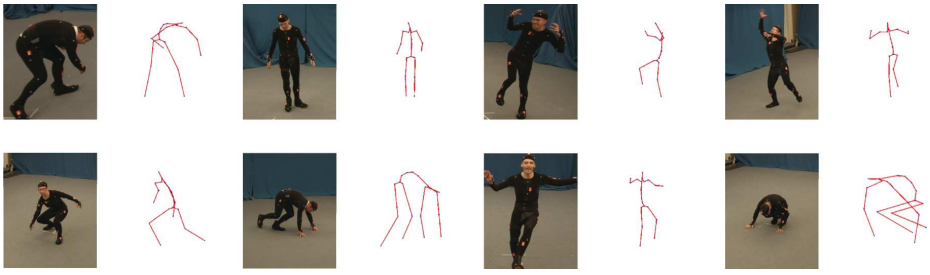

Fig. 5. Representative visual results from *TotalCapture* showing 3D pose estimation (×2 up-scaling). See Tbl. 4 for quantitative results.

is performed using the sequences of ROM1,2,3; Walking1,3; Freestyle1,2 and Acting1,2 on subjects 1, 2 and 3. We compare the pose estimation error for a number of upscale models; x1, x2, and x4 upscaling of the input PVH. Thus at the largest upscaling the PVH volume vector is $v \in \mathbb{R}^{128\times128\times128}$. Tbl. 4 shows the results of the different upscaling models against the previous state of the art for the dataset.

All three learnt upscaling models reduce the mean error of the joints by over 50% compared to previously published works for this dataset, with the error for some subjects sequences being reduced by an order of magnitude. Figure 5 provides some examples of the actions performed by the subjects and the excellent ability of the approach to estimating the pose.

Also, the table presents the *AutoEnc-FrontHalf* results, this shows initial convolutional encoder, without the decoder loss constraints. It provides a far higher error measure, indicating the importance of the dual loss constraining the skeleton pose space during training and inference. It is possible to examine the per frame error for subject 3, sequence Acting3, in Fig 6. This figure shows how consistently low the error is across the full sequence. despite a number of challenging poses being performed by the actor. There are a few error peaks, especially at the centre point, and these are generally caused by a failure in the background segmentation from which the input PVH is generated, resulting in, for example, missing or weakly defined limb extremities. This data is under-represented within the training data, however, otherwise the error is low. Use of the symmetrical network and dual loss has provided a large reduction in joint error for the skeleton it is also possible to upscale the initially very coarse and small volume at up to 4× times. Figure 7 displays the initial volume estimate, the 4x upscaled volume and the skeleton estimate for 1x, 2x and 4x for a selection of example frames on the *TotalCapture* dataset. The pose estimate for each upscaled model (1×,2× and 4×) is nearly identical as born out by the results previously presented in Tbl. 4. however, the volume enhancement from the 4× upscaling is impressive allowing for greater details to be hallucinated without noise or degeneration. Tbl. 5 compares the input and output PVH volumes against a groundtruth high resolution volume generated directly from the camera views.

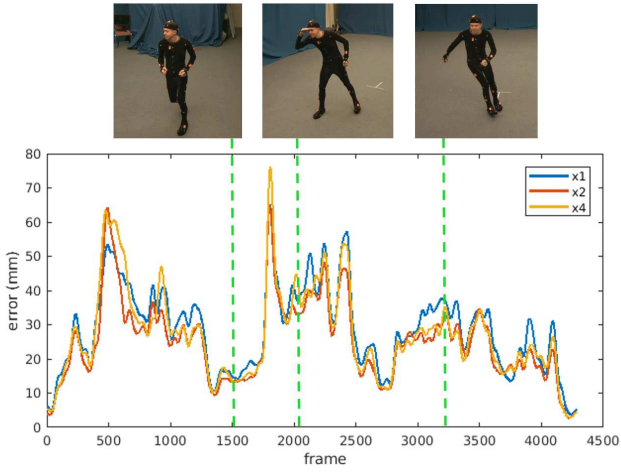

Fig. 6. Per frame skeletal error millimetres (mm) per joint on subject *S3 A3* in the *TotalCapture* test sequence.

The input volume is a naive tricubic upsampled volume and the error metric is MSE. The table shows that an order of magnitude improvement occurs using

Approach	SeenSubjects(S1,2,3)			UnseenSubjects(S4,5)			Mean
	W2	FS3	A3	W2	FS3	A3	
AutoEnc-x2 input	9.27	10.14	9.65	9.80	10.66	10.21	9.88
AutoEnc-x2 output	0.34	0.37	0.34	0.40	0.46	0.39	0.37
AutoEnc-x4 input	9.83	10.83	10.19	10.64	11.45	11.03	10.56
AutoEnc-x4 output	0.50	0.55	0.50	0.58	0.68	0.59	0.56

Table 5. Accuracy of generated volumes compared to tri-cubic upsampled input, over TotalCapture dataset. Expressed as mean voxel squared error $\times 10^{-3}$ from ground truth high resolution volume

our proposed method against a naive tricubic up-sampling method. Comparing the x2 and x4 outputs, the MSE increases only slightly despite the generative doubling of the actor volume. An illustration of the upscaling performance is shown in Figure 8, where the input and output volumes at up to x4 upscaling are shown for the TotalCapture dataset.

Despite the initial block low-res PVH, we are able to accurately generate a hi-res PVHs at up to 4 times the size, that compare favorably to a natively generated (i. e. $\mathbb{R}^{128\times128\times128}$) PVH. we are able to maintain extremity and no phantom volumes are formed in the upscaling process. Figure 9 shows the per frame MSE over a sequence, for x2 and x4 upscaling. There is little difference between the two scales despite the greatly increased volume. Table 6 shows the training and inference times (the latter near real-time) of our approach.

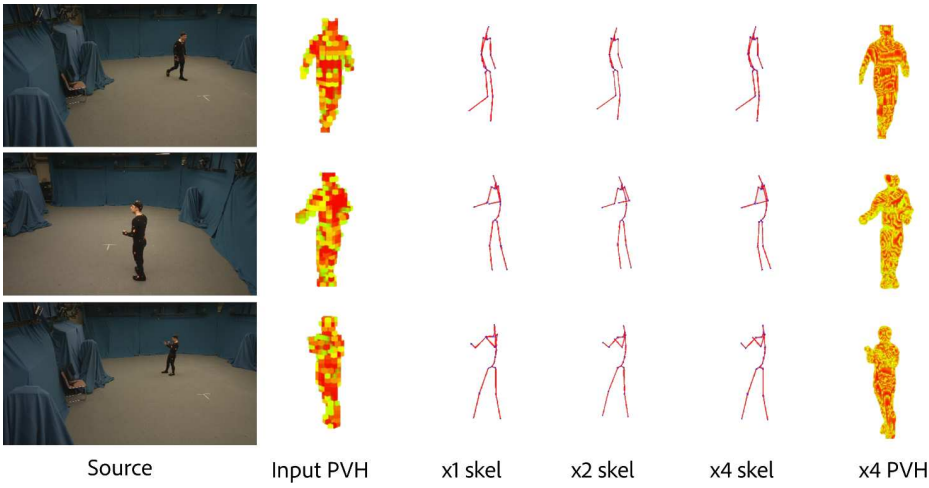

Fig. 7. Results illustrating the $\times 1, \times 2, \times 4$ upscaled volume for a representative coarse PVH alongside upscaling inferred skeletons.

4.3 Outdoor footage evaluation

To further demonstrate the flexibility of our upscaling and pose estimation approach, we test on a recent challenging dataset, *TotalCaptureOutdoor* [24]. This is a multi-view video dataset captured outdoors in challenging uncontrolled conditions with a moving and varying background of trees and differing illumination. There are 6 video cameras placed in a 120° arrangement around the subject, with a large 10x10m capture volume used. This large capture volume means the subjects are small in the scene as shown in Figure 10 below. For this dataset there are no released mattes, therefore we background subtraction was performed as a per pixel difference in HSV colour space to provide robust invariance against illumination change. There is no groundtruth annotation available for *TotalCaptureOutdoor*, however, we are present several illustrative results on two sequences: *Subject1, Freestyle, and Acting1.* Given the small size of the subjects, a traditional 3D pose estimation or volume reconstruction would be challenging. However as shown in Figure 11 we are able to use a small blocky low resolution PVH volume, that is upscaled by a factor of $\times 4$ to produce a smooth approximation of the distant subject together with an accurate estimation of their joints. Furthermore,

	Encoder Pre-train		Full Training		Inference
PVH Scale	Epochs to converge	minutes/epoch	Epochs to converge	minutes/epoch	millisec
x1	50	34	20	71	15
x2	42	32	40	58	21
x4	13	43	23	180	313

Table 6. Computational cost of model training and inference (TotalCapture dataset)

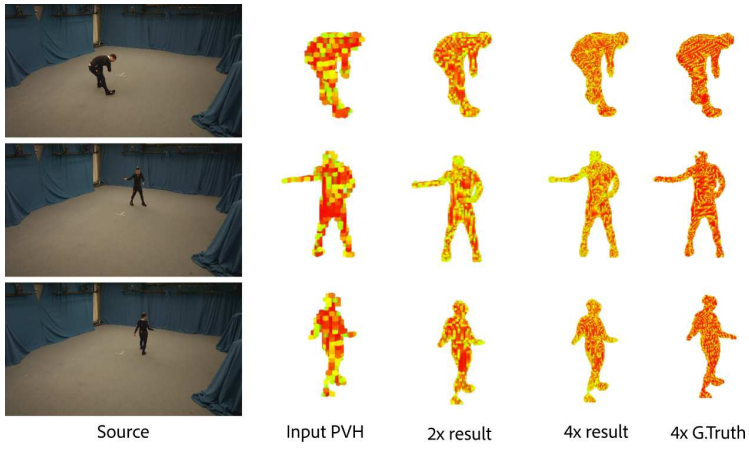

Fig. 8. Illustration of the upscaling ability of our approach on the TotalCapture dataset together with the native 128x128x128 groundtruth PVH

despite the camera being arranged in a 120° arc, we are able to simulate novel viewpoints of the upscaled full volume as shown in Figure 12, where complete 360 views are possible. This upscaling enables a future avenue of work, creating a 3D model of the upscaled volume to produce VR/AR compositions or for film/sports post-production.

5 Conclusion

We proposed a deep representation for volumetric (3D) human body shape driven by a latent encoding for the skeletal pose that can, in turn, be inferred from very coarse ($\mathbb{R}^{32\times32\times32}$) volumetric shape data. In a single inference pass our convolutional autoencoder both up-scales up the provided volumetric data (demonstrated to a factor of 4×) and predicts 3D human pose (joint positions) with greater or equal accuracy to state of the art deep human pose estimation approaches.

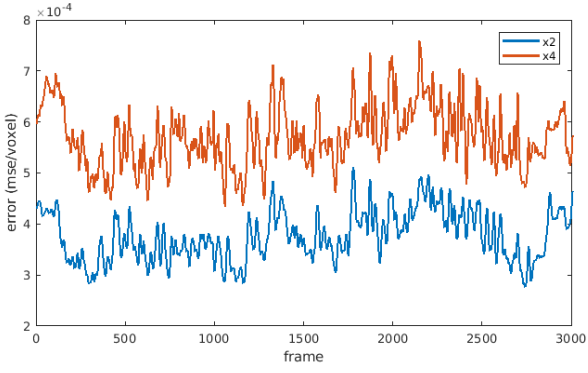

Fig. 9. Plotting volumetric reconstruction error per frame (MSE/voxel) on unseen subject *S4 A3* of the *TotalCapture* test sequence.

Fig. 10. *TotalCaptureOutdoor* dataset; red box indicates the person in the scene.

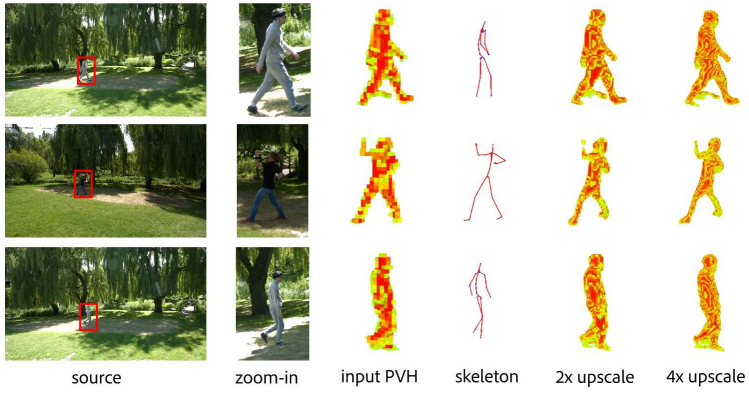

Fig. 11. Representative TotalCaptureOutdoor results showing the low-res input PVH, and resulting skeleton and upscaled volumes

Future work could explore the end-to-end integration of the LSTM to the autoencoder during training since the latter currently learns no temporal prior to aid pose or volume regression. Nevertheless, we achieve state of the art results on very low resolution volumetric input, indicating the technique has potential to enable behavioural analytics using multi-view video footage shot at a distance.

Acknowledgements

The work was supported by an EPSRC doctoral bursary and InnovateUK via the TotalCapture project, grant agreement 102685. The work was supported in part through the donation of GPU hardware by the NVidia corporation.

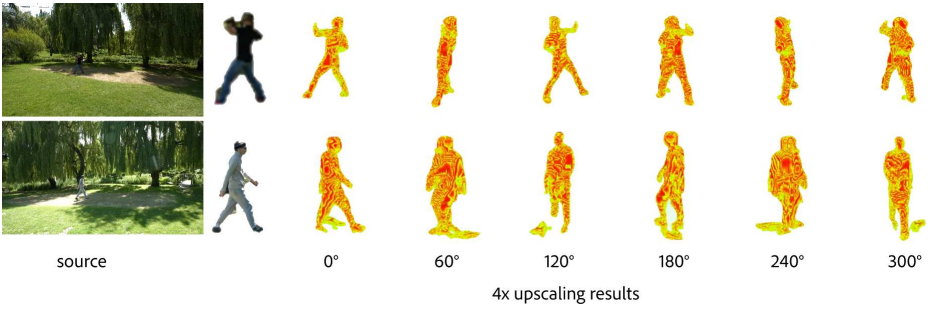

Fig. 12. Visualising the upscaled volumes from novel viewpoints. 3D reconstruction is of high quality despite the input PVH being captured from just two viewpoints.

References

1. Grauman, K., Shakhnarovich, G., Darrell, T.: A bayesian approach to image-based visual hull reconstruction. In: Proc. CVPR. (2003)
2. Fattal, R.: Image upsampling via imposed edge statistics. In: Proc. ACM SIGGRAPH. (2007)
3. Glasner, D., Bagon, S., Irani, M.: Super-resolution from a single image. In: Proc. Intl. Conf. Computer Vision (ICCV). (2009)
4. Zhu, Y., Zhang, Y., Yuille, A.L.: Single image super-resolution using deformable patches. In: Proc. Comp. Vision and Pattern Recognition (CVPR). (2014) 2917–2924
5. Freeman, W.T., Jones, T.R., Pasztor, E.C.: Example based super-resolution. IEEE Comp. Graphics and Applications **2** (2002)
6. Aydin, V., Foroosh, H.: Volumetric super-resolution of multispectral data. In: Corr. arXiv:1705.05745v1. (2017)
7. Schmidt, U., Jancsary, J., Nowozin, S., Roth, S., Rother, C.: Cascades of regression tree fields for image restoration. IEEE Trans. Pattern Anal. Machine Intelligence **38**(4) (2016) 677–689
8. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.: Extracting and composing robust features with denoising autoencoders. In: Proceedings Intl. Conf. Machine Learning (ICML). (2008) 1096–1103
9. Hayat, K.: Super-resolution via deep learning. CoRR **abs/1706.09077** (2017)
10. Xie, J., Xu, L., Chen, E.: Image denoising and inpainting with deep neural networks. In: Proc. Neural Inf. Processing Systems (NIPS). (2012) 350–358
11. Wang, Z., Liu, D., Yang, J., Han, W., Huang, T.S.: Deep networks for image super-resolution with sparse prior. In: Proc. Intl. Conf. Computer Vision (ICCV). (2015) 370–378
12. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Machine Intelligence **38**(2) (2016) 295–307
13. Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, A., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient subpixel convolutional neural network. In: Proc. Comp. Vision and Pattern Recognition (CVPR). (2016)
14. Jain, V., Seung, H.: Natural image denoising with convolutional networks. In: Proc. Neural Inf. Processing Systems (NIPS). (2008) 769–776
15. Abrahamsson, S., Blom, H., Jans, D.: Multifocus structured illumination microscopy for fast volumetric super-resolution imaging. Biomedical Optics Express **8**(9) (2017) 4135–4140
16. Brock, A., Lim, T., Ritchie, J.M., Weston, N.J.: Generative and discriminative voxel modeling with convolutional neural networks
17. Felzenszwalb, P., Huttenlocher, D.: Pictorial structures for object detection. Intl. Journal on Computer Vision **61** (2003)
18. Andriluka, M., Roth, S., Schiele, B.: Pictoral structures revisited: People detection and articulated pose estimation. In: Proc. Computer Vision and Pattern Recognition. (2009)
19. Lan, X., Huttenlocher, D.: Beyond trees: common-factor model for 2d human pose recovery. In: Proc. Intl. Conf. on Computer Vision. Volume 1. (2005) 470–477
20. Jiang, H.: Human pose estimation using consistent max-covering. In: Intl. Conf. on Computer Vision. (2009)

21. Huang, P., Tejera, M., Collomosse, J., Hilton, A.: Hybrid skeletal-surface motion graphs for character animation from 4d performance capture. ACM Transactions on Graphics (ToG) (2015)
22. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM Transactions on Graphics (TOG) **34**(6) (2015) 248
23. von Marcard, T., Rosenhahn, B., Black, M., Pons-Moll, G.: Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. Computer Graphics Forum 36(2), Proceedings of the 38th Annual Conference of the European Association for Computer Graphics (Eurographics) (2017)
24. Malleson, C., Gilbert, A., Trumble, M., Collomosse, J., Hilton, A.: Real-time full-body motion capture from video and imus. In: 3DV. (2017)
25. Tan, J., Budvytis, I., Cipolla, R.: Indirect deep structured learning for 3d human body shape and pose prediction. In: BMVC. (2017)
26. Toshev, A., Szegedy, C.: Deep pose: Human pose estimation via deep neural networks. In: Proc. CVPR. (2014)
27. Park, D., Ramanan, D.: Articulated pose estimation with tiny synthetic videos. In: Proc. CHA-LEARN Workshop on Looking at People. (2015)
28. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. ECCV'16 (2016)
29. Sanzari, M., Ntouskos, V., Pirri, F.: Bayesian image based 3d pose estimation. In: European Conference on Computer Vision, Springer (2016) 566–582
30. Zhou, X., Zhu, M., Leonardos, S., Derpanis, K.G., Daniilidis, K.: Sparseness meets deepness: 3d human pose estimation from monocular video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 4966–4975
31. Trumble, M., Gilbert, A., Hilton, A., Collomosse, J.: Deep convolutional networks for marker-less human pose estimation from multiple views. In: Proceedings of the 13th European Conference on Visual Media Production (CVMP 2016). CVMP 2016 (2016)
32. Trumble, M., Gilbert, A., Malleson, C., Hilton, A., Collomosse, J.: Total capture: 3d human pose estimation fusing video and inertial sensors. In: Proceedings of 28th British Machine Vision Conference. 1–13
33. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3D human pose. In: CVPR. (2017)
34. Tekin, B., Katircioglu, I., Salzmann, M., Lepetit, V., Fua, P.: Structured prediction of 3d human pose with deep neural networks. In: BMVC. (2016)
35. Zeiler, M.D.: Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701 (2012)
36. Hochreiter, S., Schmidhuber, J.: Long short-term memory. In: Neural computation. Volume 9., MIT Press (1997) 1735–1780
37. Sak, H., Senior, A., Beaufays, F.: Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In: Fifteenth Annual Conference of the International Speech Communication Association. (2014)
38. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 2625–2634
39. Luo, Y., Ren, J., Wang, Z., Sun, W., Pan, J., Liu, J., Pang, J., Lin, L.: Lstm pose machines. arXiv preprint arXiv:1712.06316 (2017)

40. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: 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) (jul 2014) 1325–1339
41. Malleson, C., Volino, M., Gilbert, A., Trumble, M., Collomosse, J., Hilton, A.: Real-time full-body motion capture from video and imus. In: 2017 Fifth International Conference on 3D Vision (3DV). (2017)
42. Li, S., Zhang, W., Chan, A.B.: Maximum-margin structured learning with deep networks for 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 2848–2856
43. Tekin, B., Márquez-Neila, P., Salzmann, M., Fua, P.: Fusing 2d uncertainty and 3d cues for monocular body pose estimation. arXiv preprint arXiv:1611.05708 (2016)
44. Tome, D., Russell, C., Agapito, L.: Lifting from the deep: Convolutional 3d pose estimation from a single image. arXiv preprint arXiv:1701.00295 (2017)
45. Mude Lin, Liang Lin, X.L.K.W., Cheng, H.: Recurrent 3d pose sequence machines. In: CVPR. (2017)
46. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. ICCV (2017)