Semantic Estimation of 3D Body Shape and Pose using Minimal Cameras

Andrew Gilbert · Matthew Trumble · Adrian Hilton · John Collomosse

Abstract We present an approach to accurately estimate high fidelity markerless 3D pose and volumetric reconstruction of human performance using only a small set of camera views (≈ 2). Our method utilises a dual loss in a generative adversarial network that can yield improved performance in both reconstruction and pose estimate error. We use a deep prior implicitly learnt by the network trained over a dataset of view-ablated multi-view video footage of a wide range of subjects and actions. Uniquely we use a multi-channel symmetric 3D convolutional encoder-decoder with a dual loss to enforce the learning of a latent embedding that enforces skeletal joint positions and a deep volumetric reconstruction of the performer. An Extensive evaluation is performed with state of the art performance reported on three datasets; Human 3.6M [22], TotalCapture [10] and TotalCaptureOutdoor [29]. The method opens the possibility of high-end volumetric and pose performance capture in on-set and prosumer scenarios where time or cost prohibit a high witness camera count.

1 Introduction

Performance capture is used extensively within biomechanics and the creative industries for the capture and analysis of human motion. Commercial technologies generally focus upon skeletal pose estimation and often require special (e.g. infra-red retro-reflective) markers to be worn by the subject. This work aims to jointly perform real-time video-based performance capture, able to accurately estimate both skeletal and volumetric information of a subject. However, without the need to instrument the subject with markers and using a minimal set (in general only two) of wide baseline cameras. The motivation for this comes from considering real-world scenarios away from a perfect studio environment, where only a couple of camera views are used to capture the subject as in Fig. 1. Where limitations on camera cost or placement occur such as in applications like security or sports footage or prosumer scenarios. This work proposes to use a deeply learnt before re-
channics, where the analysis of human performance data has been used to inform diagnosis and training strategy. However, the past decade has seen applications of mo-cap broaden to include performance capture, e.g. to add realism and reduce the cost of character animation in the creative industries. However, existing commercial solutions (e.g. Vicon, OptiTrack) are typically reliant upon specialist camera equipment such as active or retro-

reflective infra-red markers, stereo-triangulation depth sensors and time-of-flight cameras. While research approaches are highly effective in 2D pose estimation or through the inclusion of addition sensors such as IMUs, or the requirement of many cameras. These place restrictions on the capture environment, such as prohibiting or limiting outdoor shoots, as well as restricting the size of the capture volume.

Therefore, we propose to incorporate the use of a joint deeply learnt prior through a dual loss composed of pose and volumetric reconstruction enabling minimisation of the number of camera views required at acquisition. To further enforce accurate reconstruction and therefore, pose estimation, a generative adversarial network is used to improve realism. Specifically, we propose a convolutional encoder-decoder architecture, commonly applied to visual content for de-noising and up-scaling (super-resolution). Where the latent bottleneck is partially constrained to estimate the 3D skeletal pose and partially unimpeded to enhance the fidelity of volumetric reconstructions derived from just a few wide-baseline camera viewpoints.

We describe an encoder-decoder based generative adversarial network (GAN) with 3D convolutional stages capable of concurrently refining a probabilistic visual hull (PVH) (i.e. voxel occupancy and semantic 2D detection data derived from a small set of views) of approximately equal fidelity and 3D pose accuracy to that obtainable from the same performance captured with significantly higher (double or more) camera viewpoints. The GAN encourages the refinement of the volumetric solution to enable it to be perceptually indistinguishable from real high-fidelity reconstructions restoring fine detail such as hands and legs. Our approach extends use scenarios for performance capture to stages with low camera counts, prosumer scenarios where cost similarly limits the number of available camera views, or settings where volumetric capture is not possible due to restrictions on camera placement and cost such as sports events.

This work is based on the approach presented in however, it is greatly enhanced and improved with several core additional contributions.

- A generative adversarial network is employed to ensure the resulting high-fidelity volumetric reconstruction proxy is realistic and accurate.
- Greatly improved results over previously published works and an extensive investigation and analysis of the approach.

2 Related Work

Our work spans two classic computer vision research fields: super-resolution (SR) and human pose estimation.

Super-resolution: The classical solution to image restoration and super-resolution was to combine multiple data sources (e.g. multiple images obtained at sub-pixel misalignments, or use self-similar patches within a single image), and then incorporate these within a regularisation constraint e.g. total variation. Microscopy has applied super-resolution for volumetric data via depth of field, and through multispectral sensing data via sparse coding a machine learning-based super-resolution approach that learns the visual characteristics of the supplied training images, then applies the learnt model within an optimisation framework to enhance detail. More recently, as with all computer vision domains convolutional neural network (CNN) autoencoders have been applied to image and video up-scaling. While symmetric autoencoders have effectively learnt an image transformation between clean and synthetically noisy images. Similarly, Dong trained end-to-end networks to model image up-scaling or super-resolution.

Human Pose Estimating: There are two distinct categories of Human pose estimation; bottom-up data-driven and top-down, fitting an articulated limb kinematic model to the source data. In general, top-down 2D pose estimation fit a previously defined articulated limb model to data incorporating kinematics into the optimisation to bias toward possible configurations. The model can be user-defined or learnt through a data defined model such as the SMPL Body Model. Lan define a model and consider the conditional independence of parts; however, inter-limb dependencies (e.g. symmetry) are not considered. Jiang considers a more global treatment using linear relaxation but performs well only on uncluttered scenes.

Bottom-up pose estimation is driven by image parsing to isolate components, Srivivasan used graph-cuts to parse a subset of salient shapes from an image and group these into a model of a person. Ren recursively splits Canny edge contours into segments, classifying each as a putative body part using cues such as parallelism. Ren also used Bag of Visual Words for implicit pose estimation as part of a pose similarity system for dance video retrieval. More recently studies have begun to leverage the power of convolutional neural networks, following in the wake of the eye-opening results...
Semantic Estimation of 3D Body Shape and Pose using Minimal Cameras

Fig. 2 Network architecture, it takes two wide baseline camera views and produces a low fidelity geometric proxy, this proxy is passed through a decoder-encoder to produce a 3D pose estimate and a high-fidelity geometric proxy. The geometric proxy is used as an input to a Discriminator network to improve the quality of the proxy.

of Krizhevsky [25] on image recognition. In DeepPose, Toshev [47] used a cascade of convolutional neural networks to estimate 2D pose in images. Descriptors learnt by a CNN have also been used in 2D pose estimation from very low-resolution images [33]. Elhayek [8] used MVV with a Convnet to produce 2D pose estimations while Rhodin [38] minimised the edge energy inspired by volume ray casting to deduce the 3D pose.

Estimating 3D pose from 2D joints More recently given the success and accuracy of 2D joint estimation [4], several works lift 2D detections to 3D using learning or geometric reasoning, aiming to recover the missing depth dimension in the images. Sanzari [40] estimates the location of 2D joints, before predicting 3D pose using appearance and probable 3D pose of the discovered parts with a hierarchical Bayesian model. While Zhou [55] integrates 2D, 3D and temporal information to account for uncertainties in the data. The challenge of estimating 3D human pose from MVV is currently less explored, generally casting 3D pose estimation as a coordinate regression task, with the target output being the spatial x, y, z coordinates of a joint with respect to a known root node such as the pelvis. Trumble [49] used a flattened MVV based spherical histogram with a 2D convnet to estimate pose. While Pavlakos [34] used a simple volumetric representation in a 3D convnet for pose estimation and Wei [52] performed related work in aligning pairs of joints to estimate 3D human pose. Differently, Huang [21] constructed a 4-D mesh of the subject from video reconstruction to estimate the 3D pose.

Using Temporal Information Since detecting pose for each frame individually leads to incoherent and jittery predictions over a sequence, many approaches exploit temporal information. Andriluka [2] used tracking-by-detection to associate 2D poses detected in each frame individually and used them to retrieve 3D pose. While Tekin [45] used a CNN to first align bounding boxes of successive frames so that the person in the image is always at the centre of the box and then extracted 3D HOG features over the spatiotemporal volume from which they regress the 3D pose of the central frame. Lin [32] performed a multi-stage sequential refinement using LSTMs [19] to predict 3D pose sequences using previously predicted 2D pose representations and 3D pose. While Hossain [20] learns the temporal context of a sequence using a form of sequence-to-sequence network.

Our work shares the dual goals of 3D pose estimation from MVV and the high-level goal of learning deep models for detail enhancement. However, we utilise volumetric (PVH) data and seek not to up-scale (increase resolution) as in SR, but instead, conjointly estimate the 3D pose constrained with enhanced detail within a voxel grid to simulate the benefit of having additional viewpoints available during the formation of the PVH and pose estimation.

3 Joint minimal camera Pose and Volume reconstruction

The goal of our method is to learn a generative model that accepts a coarse poor quality volumetric proxy formed from a low number of wide baseline camera views of a subject. Then in a single inference step, estimate both the skeletal joint positions and refine a higher fidelity volumetric reconstruction. A joint loss between both outputs is used within a generative adversarial network to ensure realistic reconstruction.
Our process for refining the poor quality volume reconstruction echoes the two-stage process employed in traditional image denoising, first, a pre-processing step [16] reconstructs a coarse Probabilistic Visual Hull (PVH) proxy using a limited number of cameras (Sec. 3.2), constructed from occupancy and semantic 2D joint estimates (Sec. 3.1) and will contain phantom limbs and additional false positive voxels. Secondly, a 3D convolutional encoder-decoder (Sec. 3.3) generative adversarial network (GAN) (Sec. 3.4) learns a deep representation of body shape and the skeletal pose encoding with a dual loss. The feature representation of the PVH (akin to a low-fidelity image in super-resolution pipelines), is deeply encoded via a series of convolution layers, embedding the skeletal joint positions in a latent or hidden layer, concatenating the joint estimates with an additional unconstrained feature representation. This latent space enables non-linear mapping decoding to a high fidelity PVH (akin to the high-fidelity image), while the 3D joint estimations are fed to LSTM layers to enforce the temporal consistency of the 3D joints (Sec. 3.5). We also describe the data augmentation and methodology for training the 3D convolutional network (Sec. 4).

3.1 Visual Features

To estimate the pose, we propose to use visual features that form a 3D voxel probability from two distinct modes created from RGB images, a 2D foreground occupancy matte and 2D semantic joint detections. The probabilistic occupancy provides a low fidelity shape-based feature, relatively invariant to appearance and clothing, that complements a semantic contextual 2D joint estimate that provides internal feature description. To compute the matte, the difference between the current frame observed by a limited number of cameras and the previous image or the previous stage returned pixel-wise belief map. At each stage and for each joint label the algorithm returns dense per pixel belief maps \( n^j_{z}(x,y) \), which provides the confidence of a joint centre for any given pixel \((x,y)\). The per joint belief maps are maximised over the confidence of all possible joint labels to produce a single label per pixel image \( M(x,y) \).

\[
M(x,y) = \arg \max_j n^j_{z}(x,y)
\]

Eqn. (1)

Fig. 3 illustrates the 2D occupancy and semantic joint labels for an example frame, for this complex pose the occupancy shape is detected well, however, there is ambiguity over the 2D pose. There is a failure for a couple of the 2D semantic joint estimates due to uncertainties with the left arm and head. However, by jointly using both modes and learning to model and encode their response as a probabilistic visual hull, an accurate 3D pose estimate and shape proxy can be identified.

![Image](Image333x529 to 500x687)

**Fig. 3** An example of the foreground occupancy and 2D semantic labels converted into probably for PVH construction.

3.2 Volumetric Representation

There exist several methods to estimate 3D pose; through multiple separate 2D views \([35,39]\) that require many cameras in the scene or by inferring 3D from a single 2D view \([40,43]\), which can fail for complex poses occluded by the single-camera view. However, by taking inspiration from super-resolution work, we propose a learn a generative approach that uses a minimal number of camera views and an inherent poor input to learn a complex mapping to a multi-view 3D pose previously learnt from many camera views. Thus, learning to resolve complex ambiguities and occlusions present in individual 2D images. To construct our data representation consisting of a volume voxel, we use a multi-channel based probabilistic visual hull (PVH).

We assume a capture volume observed by a limited number \( C \) of camera views \( \{1,...,C\} \) for which extrinsic parameters \( \{R_c,COP_c\} \) (camera orientation and focal point) and intrinsic parameters \( \{f_c,o_c^x,o_c^y\} \) (focal length, and 2D optical centre) are known. An external process, e.g. a person tracker, isolates the bounding sub-volume \( X_I \in V \) corresponding to, and centred upon, a single subject of interest, and which is finely decimated into voxels \( V^i_L = \{v^i_x,v^i_y,v^i_z\} \) for \( i = 1,...,|V_L| \); each voxel is approximately 5mm\(^3\) in size. Each voxel \( v^i \in V_L \) projects to coordinates \((x[v^i],y[v^i])\) in each camera view \( c \) derived in homogeneous form via pin-hole projection:

\[
\begin{bmatrix}
\alpha x[v^i] \\
\alpha y[v^i] \\
\alpha
\end{bmatrix}
= \begin{bmatrix}
  f_c & 0 & o_c^x & 0 \\
  0 & f_c & o_c^y & 0 \\
  0 & 0 & 1 & 0
\end{bmatrix}
\begin{bmatrix}
  -R_c^{-1}T_c \\
  v^i_x \\
  v^i_y \\
  v^i_z
\end{bmatrix}
\]

Eqn. (2)
tation likelihoods obtained, through background subtraction or semantic 2D joint estimates, the point \((x_c, y_c)\) is the point within \(I_c\) to which \(V_{L}^i\) projects in a given view:

\[
x[V_{L}^i] = \frac{f_{x} v_{x}^{i}}{v_{z}^{i}} + \sigma_x^{i} \quad \text{and} \quad y[V_{L}^i] = \frac{f_{y} v_{y}^{i}}{v_{z}^{i}} + \sigma_y^{i},
\]

(3)

\[
[v_{x}^{i}, v_{y}^{i}, v_{z}^{i}] = COPC - R_{c}^{-1}V_{L}^i.
\]

(4)

The probability of the voxel being part of the performer in a given view \(c\) is:

\[
p(V_{L}^i|c) = I_{c}(x[V_{L}^i], y[V_{L}^i], \phi).
\]

(5)

The overall probably of occupancy for a given voxel \(p(V_{L}^i, \phi)\) is:

\[
p(V_{L}^i, \phi) = \prod_{i=1}^{C} \frac{1}{(1 + e^{p(V_{L}^i|c)})}.
\]

(6)

We are then able compute \(p(V_{L}^i)\) for all voxels to create the PVH for volume \(V_L\).

\[
\sum_{i} \sum_{j \in \Phi} p(v_i, \phi_j)
\]

(7)

### 3.3 Dual Loss Convolutional Volumetric Network

At their simplest, an encoder-decoder neural network architecture learns an encoding from an input signal domain by training the network to reconstruct the input through a bottleneck layer of reduced dimensionality (the latent embedded space). This layer is a concatenation of the 3D pose estimates and a vector with no direct constraint and as such is referred to as the hidden or latent layer. By using multiple 2D views, our result can generate a realistic 3D representation and pose of the human body, thus able to avoid ambiguities and occlusions present in independent, individual 2D images.

We propose to learn a deep representation or output given an input tensor \(V_L\) where \(V_L \in \mathbb{R}^{X \times Y \times Z \times \phi}\), where each dimension encodes the probability of volume occupancy \(p(X, Y, Z)\) derived from a PVH obtained using a low camera count (Eq.6) from channels (\(\phi\)); foreground occupancy and semantic 2D joint estimates. We wish to train a deep representation to solve the prediction problem \(V_H = F(V_L)\) for similarly encoded tensor \(V_H \in \mathbb{R}^{W \times H \times D \times \phi}\) derived from a higher fidelity PVH of identical dimension obtained using a higher camera count. Where \(W, H, D, \phi\) are the width, height, depth and channel of the performance capture volume respectively. Function \(F\) is learnt using a CNN, specifically a convolutional Sec. 3.3 consisting of successive three-dimensional (3D) alternate convolutional filtering operations and down- or up-sampling with nonlinear activation layers for a similarly encoded output tensor \(V_H\), where

\[
V_H = F(V_L) = D(E(V_L))
\]

(8)

for the learnt encoder \((E)\) and decoder \((D)\) functions. The encoder yields a latent feature representation via a series of 3D convolutions. Each convolutional layer is followed by batch normalisation and a ReLU in the Generator and convolutional strides for a layer in both the encoder and decoder. The encoder enforces \(J(V_L) = E(V_H)\) where \(J(V_L)\) is a concatenation of the skeletal pose vector corresponding to the input PVH; specifically a 78-D vector concatenation of 26 3D Cartesian joint coordinates in \(x, y, z\) to generate the pose estimate and an additional latent embedding of size \(e\) (in general \(e = 200\)). The decoder half of the network inverts this process to output tensor \(V_H\) matching the input resolution but with higher fidelity content.

Fig. 4 illustrates the network architecture which incorporates two skip connections bypassing the network bottleneck to allow the output from a convolutional layer in the encoder to feed into the corresponding deconvolution layer in the decoder. Combining the activations from the preceding layer in the main network and skip connection data via mean average, the use of mean average combination instead of element-wise addition or concatenation is analysed later in section 5.8.

Fig. 2 illustrates our symmetric architecture with skip connections bridging hourglass encoder-decoder stages, the full network parameters are:

- \(n_E = [64, 64, 128, 128, 256],\)
- \(n_D = [256, 128, 128, 64, 64],\)
- \(k_E = [3, 3, 3, 3, 3],\)
- \(k_D = [3, 3, 3, 3, 3],\)
- \(k_s = [0, 1, 0, 1, 0],\)

where \(k[n]i\) indicates the kernel size and \(n[i]\) is the num-
The encoder-decoder model described in the section volumetric representation given an initial poor fidelity voxels through the two losses $L_J$ parts, while simultaneously producing the 3D joint position volume with a discriminator loss, using the theory introduced by Goodfellow [15] defined as a game between two competing networks: the Discriminator $D$ and train both networks in an adversarial setup. The Discriminator having the objective of maximising the chance of recognising real PVH volumes as real and generated PVH volumes as fake, i.e. the maximum likelihood of the observed data. The goal of the Generator is to fool the Discriminator by generating perceptually convincing samples indistinguishable from a real one. The game between the Generator $G$ and Discriminator $D$ is the minimax objective:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim P_r}[\log(D(x))] + \mathbb{E}_{\tilde{x} \sim P_g}[\log(1 - D(\tilde{x}))]$$

where $P_r$ is the (real) data distribution and $P_g$ is the (generated) model distribution, defined by $\tilde{x} = G(z), z \sim P(z)$, where the input $z$ is a sample from a simple noise distribution. Once both objective functions are defined, they are learnt jointly by the alternating gradient descent. Initially, the decoder part of the Generator is pre-trained to learn a 3D pose estimate without the constraint of the 3D proxy, to produce the initial latent embedding. Once converged; alternatively, we train the full network, initially, we train the Generator models parameters for a single iteration and fix the Discriminators parameters. Then a single iteration of gradient descent on the Discriminator using the real and the generated images is performed. Then the Discriminator is fixed, and train the Generator for further iteration. Both networks are trained in alternating steps until the Generator produces good quality volume reconstructions using the dual loss found in Eq. 4 An alternate training period for the network parts produced the most stable training process. The Discriminator network is shown in Fig. 2, it consists of 3D convolutional layers, which are followed by each time by batch normalisation and leaky ReLu activations.

### 3.4 Generative Adversarial Network Model

The encoder-decoder model described in the section above with the dual volume and joint pose loss can produce reasonable results. However, we propose to constrain and improve the reconstruction quality of the decoder output of the 3D occupancy volume and the pose estimation by employing a generative adversarial network. Enforcing the learning of a realistic 3D occupancy volume with a discriminator loss, using the theory introduced by Goodfellow [15] defined as a game between two competing networks: the Discriminator and the Generator.

The goal of the generative adversarial network is to recover a sharp, high-quality PVH volume $V_H$, given a poor low-quality volume $V_L$ with possible phantom parts, while simultaneously producing the 3D joint positions $J(V_L)$. Where the improved volume is estimated by the encoder model from section 3.3, which we refer to as the Generator $G$. Also, during the training phase, we introduce the critic network, the Discriminator $D$ and train both networks in an adversarial setup. The Discriminator having the objective of maximising the chance of recognising real PVH volumes as real and generated PVH volumes as fake. i.e. the maximum likelihood of the observed data. The goal of the Generator is to fool the Discriminator by generating perceptually convincing samples indistinguishable from a real one. The game between the Generator $G$ and Discriminator $D$ is the minimax objective:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim P_r}[\log(D(x))] + \mathbb{E}_{\tilde{x} \sim P_g}[\log(1 - D(\tilde{x}))]$$

where $P_r$ is the (real) data distribution and $P_g$ is the (generated) model distribution, defined by $\tilde{x} = G(z), z \sim P(z)$, where the input $z$ is a sample from a simple noise distribution. Once both objective functions are defined, they are learnt jointly by the alternating gradient descent. Initially, the decoder part of the Generator is pre-trained to learn a 3D pose estimate without the constraint of the 3D proxy, to produce the initial latent embedding. Once converged; alternatively, we train the full network, initially, we train the Generator models parameters for a single iteration and fix the Discriminators parameters. Then a single iteration of gradient descent on the Discriminator using the real and the generated images is performed. Then the Discriminator is fixed, and train the Generator for further iteration. Both networks are trained in alternating steps until the Generator produces good quality volume reconstructions using the dual loss found in Eq. 4 An alternate training period for the network parts produced the most stable training process. The Discriminator network is shown in Fig. 2, it consists of 3D convolutional layers, which are followed by each time by batch normalisation and leaky ReLu activations.

**Discriminator**

| Conv3D 3x3x3x128 stride2 |
| Conv3D 3x3x3x256 |
| FC |

**Fig. 5** The critic Discriminator network on the volume reconstruction.

3.4.1 Skip Connections

Deeper networks in image restoration tasks can suffer from performance degradation. As given the increased
number of convolutional layers, finer image details can be lost or corrupted, as given a compact latent feature abstraction. The recovery of the image detail is an under-determined problem, exasperated by the need to reconstruct the additional dimension in volumetric data. In the spirit of highway [43] and deep residual networks [18], we add skip connections between two corresponding convolutional and deconvolutional layers, as shown in Fig. 2. Skip connections are an architectural feature used in deeper reconstructive networks that provide a bridge across the latent layer, directly connecting corresponding encoder and decoder layers. Allowing intermediate stages of the encoder to transmit directly to latter stages of the decoder can aid the reconstruction of high-frequency detail and mitigate the vanishing gradient problem of many-layered networks by providing a new direct route for the error gradient to back-propagate to earlier layers. Our proposed skip connections differ from that proposed in recent image restoration work [18] which concern only smoother optimisation. Instead, we pass the feature activations at intervals of every two convolutional layers to their mirrored up-convolutional layers to enhance reconstruction detail. The skip connections are incorporated into the model using the mean average rather than elementwise addition. However, there is little difference in the incorporation method used on the performance of the joint accuracy and reconstruction of the volume. However, if we omit the skip connections much of the detail of the extremities such as lower arm position is poorly estimated by both the volume and 3d joints, resolving joint accuracy and reconstruction of the volume. How-ever, if we omit the skip connections much of the detail of the extremities such as lower arm position is poorly estimated by both the volume and 3d joints, resolving the mean pose and volume.

3.5 Temporal Consistency

Given the inherent temporal nature of the human pose, we enforce this consistency with additional Long Short Term Memory (LSTM) layers. These help to smooth noise in individual joint detections that would otherwise cause large estimation errors. The latent vector from the encoder $J(V_{L_t}) = E(V_{L_t})$ at time $t$ consisting of concatenated joint spatial coordinates passed through a series of gates resulting in an output joint vector $J_o$. The aim is to learn the function that minimises the loss between the input vector and the output vector $J_o = o_t \circ \text{tanh}(c_t)$ (where 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 this implementation, the model consists of two LSTM layers both with 1024 memory cells, using a look back of $T = 5$.

4 Experiential Setup

To quantify the performance of our proposed approach we report Mean Per Joint Position Error, the mean 3D Euclidean distance between ground-truth and estimated joint positions of the 26 joints including hips, knees, ankles, neck, head, shoulders, elbows and wrists; In order to evaluate pose accuracy independently of absolute camera position and orientation, we align our estimates with the ground-truth. Aligning with the ground truth is standard practice in existing benchmarks [22,12]. Thus, in our case, the Mean Per Joint Position Error is a measure of pose accuracy independent of global position and orientation.

To generate both training and test sequences, given the temporal requirement of the LSTM, we translated a sliding window of length $T$ successively by a single frame across the sequence. Hence there is an overlap between the frames, providing additional data to train on, which is always an advantage for deep learning systems. While during test time, we initially predict the first $T$ frames of the sequence and slide the window by a stride length of 1 to predict the next frame using the previous pose $T$ estimates.

To train $F$, we use Adadelta [54] an extension of Adagrad that seeks to reduce its aggressive, radically diminishing learning rates, restricting the window of accumulated past gradients to some fixed size $\omega$. Initially, training the encoder for just the skeleton loss, purely as a pose regression task without the decoder or critic networks. The fixed encoder training is due to the large size of parameters in the network and the fully connected layer in our model that is in general unsuitable for GAN models. These trained weights initialise the encoder stage to help constrain the latent representation during the full, dual-loss network training. Then given the learnt weights as initialisation for the encoder section, we train the entire encoder/decoder network end to end constrained by the dual loss of the skeleton and volume occupancy through the GAN critic network. The encoder-decoder Generator and Discriminator network are trained alternately, with the opposing network weights fixed.

The pose term of the dual loss (Eq. 9) is scaled by a factor of $\lambda$. We found the approach insensitive to this parameter up to an order of magnitude and set $\lambda = 10^{-3}$ for all experiments. Below $10^{-5}$, the bottle-neck convergences to a semantic representation of the pose that is stable but does not resemble joint angles while above $10^{-2}$ the network will not converge. Initialising the weights of the layers by the Xavier uniform initialiser [14] and we use a mini-batch batch size of 32 and a sequence length of $T = 5$. To incorporate the temporal nature into the model, we experimented with different sequence lengths and found sequence length 3, 4, 5 and 6 generally gave similar results. We augment the data during training with a random rotation around the
We quantitatively evaluate tracking accuracy on the TotalCapture dataset. The dataset consists of 5 subjects performing several activities such as walking, acting, a range of motion sequence (ROM) and freestyle motions, which are recorded using 8 calibrated, static HD RGB cameras and 13 IMUs attached to head, sternum, waist, upper arms, lower arms, upper legs, lower legs and feet, however the IMU data is not required for our experiments. The dataset has publicly released foreground mattes that we use to compute the occupancy PVH, and we use the released RGB images to localise the semantic 2D joint estimates. Ground-truth poses are obtained using a marker-based motion capture system, with the markers are < 5mm in size and therefore invisible to the training model. All data is synchronised and operates at a framerate of 60Hz, providing ground truth poses as joint positions. We study the accuracy gain due to our method by ablating the set of camera viewpoints. We outperform the single loss approaches using an autoencoder and all data modalities \cite{30,50} and our proposed approach with the dual loss GAN \cite{4}.

5 Evaluation and Discussion

To quantify the improvement in both the upscaling of low-resolution volumetric representations and human pose estimation, performing quantitative evaluation over three public multi-view video datasets of human actions. 3D human pose is evaluated for Human 3.6M \cite{22}, and the performance of both the skeleton estimation and volume reconstruction is evaluated in the TotalCapture \cite{50} and TotalCaptureOutdoor \cite{29} datasets.

5.1 TotalCapture Evaluation

We quantitatively evaluate tracking accuracy on the TotalCapture dataset. The dataset consists of 5 subjects performing several activities such as walking, acting, a range of motion sequence (ROM) and freestyle motions, which are recorded using 8 calibrated, static HD RGB cameras and 13 IMUs attached to head, sternum, waist, upper arms, lower arms, upper legs, lower legs and feet. The PVH at \( C = 8 \) provides the ideal 3D reconstruction proxy estimation for comparison, while \( C = \{2,4\} \) input covers at most a narrow 90° view of the scene. Before refinement, the ablated view PVH data exhibits phantom extremities and lacks fine-grained detail, particularly at \( C = 2 \) (Fig. 10). These crude volumes would be unsuitable for pose estimation or reconstruction as they do not reflect the true geometry and would cause poor defined joint estimations and severe visual misalignments when projecting camera texture onto the model. However, our method can estimate the joint positions accurately and also clean up and hallucinate a volume equivalent to one produced by the unablated \( C = 8 \) camera viewpoints. Tab. 1 quantifies the pose animation error between previous approaches using in general multiple camera views \cite{18,19,50,53} or additional data modalities \cite{50,30} and our proposed approach with only two camera views. We outperform the single loss learning-based approach introduced in the TotalCapture dataset \cite{50} by 48mm, this approach uses all eight cameras and fuses the data of 13 IMU sensors with the probabilistic visual hull. The approach of Pons \cite{30} also uses the 13 IMUs sensors and a single reference camera and achieve similar performance to us of 26mm; however, it requires that the full sequence is simultaneously trained and test partitions with respect to the subjects and sequences, the training consists of ROM1,2,3: Walking1,3, Freestyle1,2 and Acting1,2 on subjects 1,2 and 3.

The PVH at \( C = 8 \) provides the ideal 3D reconstruction proxy estimation for comparison, while \( C = \{2,4\} \) input covers at most a narrow 90° view of the scene. Before refinement, the ablated view PVH data exhibits phantom extremities and lacks fine-grained detail, particularly at \( C = 2 \) (Fig. 10). These crude volumes would be unsuitable for pose estimation or reconstruction as they do not reflect the true geometry and would cause poor defined joint estimations and severe visual misalignments when projecting camera texture onto the model. However, our method can estimate the joint positions accurately and also clean up and hallucinate a volume equivalent to one produced by the unablated \( C = 8 \) camera viewpoints. Tab. 1 quantifies the pose animation error between previous approaches using in general multiple camera views \cite{18,19,50,53} or additional data modalities \cite{50,30} and our proposed approach with only two camera views. We outperform the single loss learning-based approach introduced in the TotalCapture dataset \cite{50} by 48mm, this approach uses all eight cameras and fuses the data of 13 IMU sensors with the probabilistic visual hull. The approach of Pons \cite{30} also uses the 13 IMUs sensors and a single reference camera and achieve similar performance to us of 26mm; however, it requires that the full sequence is simultaneously optimised over. We do not require sensors to be placed on the human, removing the requirement to pre-make up the subject and only require an additional camera and receive a similar low joint error. We outperform the previous dual loss approach using an autoencoder and all eight cameras \cite{48} by 14mm indicating the importance of the GAN loss and semantic 2D joint estimates.

Fig. 8 illustrates the performance of the approach qualitatively on challenging frames and also in the accompanying video (The video is available at https://youtu.be/MlWTFwkbXHo). Fig. 7 illustrates the frame-wise error across all frames of all sequences in the TotalCapture dataset, our approach can maintain a low,
with no dramatic failure frames present, with the maximum mean error of only 7cm and mean of only 2cm. The error peaks are generally caused by a simultaneous failure of both channels of the PVH, the foreground occupancy and 2D semantic joints. For example, missing or weakly defined limb extremities, and such data is under-represented within the training data, the error is otherwise consistently low. To indicate the strength of the approach in 3D, we also show a challenging pose visualised from 10 hallucinated camera angles in Fig. 6.

5.2 Ablation Study

In order to better understand the influence of the individual components and design decisions of our network we also perform an ablative analysis of the tracking accuracy for the individual contributions of the approach. The results are reported in Tab. 2. Each part of the process enables an improvement in the accuracy performance, especially the use of dual loss in the approach (AutoEncLSTM) with an error of 35.5mm, the inclusion of the 2D joint (2DJoint) estimates into the dual channel PVH further reduces this loss by around 4 mm to 31.1 average joint error. The inclusion of the Discriminator (GAN8cam) to enforce improved 3D occupancy volume result, enables the loss to be further reduced to 21mm per joint using all 8 camera views. The greater the number of cameras, the more visually realistic the input dual channel PVH is. However, it is possible to remove a large number of these camera with little or no impact on performance (GAN4cam and GAN2cam). This is despite the appearance of the input PVH being greatly degraded when using only 2 or 4 views as input, as indicated by Fig. 9. The figure also illustrates the resulting output PVH and this can be seen to be of a high-fidelity result invariant to the number of cameras used. In summary, training a model that uses a low-fidelity PVH constructed from only 2 camera views with phantom and missing voxels, will still achieve a headline performance of 21.4mm mean per joint error.
Table 2 Ablation study of the Mean per joint error (mm), for the individual components on the TotalCapture Dataset.

| Approach          | Features          | Model  | SeenSubjects(S1,2,3) | UnseenSubjects(S4,5) | Mean |
|-------------------|-------------------|--------|----------------------|----------------------|------|
|                   | Occ. 2DJoint Enc Dec LSTM GAN W2 FS3 A3 | W2 FS3 A3 |
| Encoder           | 8cam              | ✓      | -                    | -                    | 15.2 65.7 54.4 |
| EncoderLSTM       | 8cam              | -      | ✓                    | ✓                    | 13.4 49.8 24.3 |
| AutoEncLSTM       | 8cam              | -      | ✓                    | ✓                    | 10.2 123.1 88.6 |
| 2DJoint           | -                 | 8cam   | ✓                    | ✓                    | 21.2 123.1 88.6 |
| Occ+2DJoint       | 8cam              | 8cam   | ✓                    | ✓                    | 8.2 30.5 15.0  |
| GAN8cam           | 8cam              | 8cam   | ✓                    | ✓                    | 9.8 29.9 15.3  |
| GAN4cam           | 4cam              | 4cam   | ✓                    | ✓                    | 9.2 30.3 15.2  |
| GAN2cam           | 2cam              | 2cam   | ✓                    | ✓                    | 8.2 30.5 15.0 |

Table 3 Quantitative performance of volumetric reconstruction on the TotalCapture dataset using 2-4 cameras before our approach (Input) and after, versus unablated groundtruth using eight cameras (error as MSE $\times 10^{-3}$).

| Method Cams SeenSubs(S1,2,3) UnseenSubs(S4,5) Mean |
|-----------------------------------------------|-------------------------------------|
| Input                                        | Input                               |
| 2cam                                         | 29.5 23.9 25.4 27.5 25.2 24.6        |
| C1                                            | 4.43 5.34 10.06 8.71 7.71             |
| Ours                                          | 5.44 9.94 6.34 5.16 9.86 8.49 7.34    |
| C1                                            | 4.85 9.32 5.84 4.83 9.56 8.03 7.02    |

5.3 Evaluating Reconstruction Accuracy

In addition to the pose estimation, the dual loss model is also able to reconstruct the high-fidelity 3D volume for the given low fidelity PVH input. Tab. 3 quantifies the error between the unablated ($C = 8$) and the reconstructed volumes for $C = \{2, 4\}$ view PVH data, baselining these against $C = \{2, 4\}$ PVH prior to enhancement via our learnt model (input).

To measure the performance, we compute the average per-frame MSE of the probability of occupancy across each sequence. Comparing the two and four camera PVH volume before enhancement and our results indicate a reduction in MSE of around three times through our approach when using two cameras views for the input and a halving of MSE for a PVH formed from 4 cameras. It is possible to observe that $C = 4$ in a 180° arc around the subject perform slightly better than $C = 2$ neighbouring views in a 90° arc. However, the performance decrease is minimal for the significantly increased operational flexibility that a two camera deployment provides. In all cases, MSE is more than halved (up to 34% lower) using our refined PVH for a reduced number of views. Using only two cameras, we can produce an equal volume to that reconstructed from a full 360° $C = 8$ setup. We show qualitative results of using only two and four camera viewpoint to construct the volume in Fig. 10 where high-quality reconstructions are possible despite the presence of phantom limbs and large false volumes in the input PVH. In all cases, performance is slightly better when testing on seen versus unseen subjects.

5.4 Human 3.6M evaluation

We perform further quantitative and qualitative evaluation on the Human 3.6M [22] dataset. Human 3.6M, is the largest publicly available dataset for human 3D pose estimation and contains 3.6 million images of 7 different professional actors performing 15 everyday activities including walking, eating, sitting, making a phone call. Each video is captured using 4 calibrated cameras arranged in the 360° arrangement and contains 3D pose ground truth formed from a standard motion capture.
system. We follow the standard protocols of the Human3.6M dataset used in the literature, using subjects 1, 5, 6, 7, and 8 for training, and subjects 9 and 11 for testing. The error is evaluated like that reported for the TotalCapture dataset on the predicted 3D pose without any transformation. The Human3.6M PVH models are poor quality as there are only 4 cameras at body height in four corners of a studio covering a relatively large capture area. This causes phantom parts and ghosting to occur as shown in the examples of the reconstructed PVH input using $C = \{4\}$ views (4cam Input) on Human3.6M in Fig. 11. Therefore, we explore the transfer of the high fidelity 8cam trained model from the TotalCapture dataset to the 4 cam human3.6M dataset through three specified methods of training:

1. A baseline direct training of the approach using the specified Human 3.6M training data with the 4 cameras producing lower quality PVH reconstructions assuming the semantic 2D joints will compensate in part for the phantom part and ghosting that occurs to the occupancy voxels, denoted as Human3.6Model.

2. Transfer of the trained model from the TotalCapture dataset, given that the trained model on TotalCapture is able to accurately estimate the human pose using only 2-4 views to approximate pose accuracy from 8 views, we transfer our trained CNN models that improves 4 $\mapsto$ 8 views on TotalCapture without any further training, to estimate pose as if 8 cameras were used at acquisition, this is denoted as TotalCaptureModel.

3. Fine Tuning of TotalCapture on the Human3.6M dataset, results using the learnt TotalCapture 4 $\mapsto$ 8 camera model, with an additional 2 epochs of fine tuning with the Human3.6M dataset, denoted as TotalCapture+FineTune(H36M Model).

The performance of the three training methods are shown in Tab. 4 together with further recent approaches and qualitative example frames shown are shown in Fig. 12.

It can be observed that our the use of the TotalCapture trained model (TotalCaptureModel) improves the baseline training of Human 3.6M (Human3.6Model) alone by 5mm and the combined TotalCapture of fine-tuned model TotalCapture+FineTune(H36M Model) dramatically improves this performance by a further 10mm. Our network significantly improves the state-of-the-art result of Imtiaz [20] by approximately 6mm. By using the information of temporal context and semantic joint estimations, our network reduces the overall error in estimating 3D joint locations, especially on actions like phone, photo, sit and sitting down on which most previous methods did not perform well due to heavy occlusion. Our method learnt the temporal context of the sequences and predicted temporally consistent 3D poses. The results presented for the datasets are using the standard train and test split to enable direct comparison with other approaches, however in order to further demonstrate robustness to the train and test subjects we performed additional tests by performing five rounds of cross-validation using multiple pairs of different test subjects with the remaining subjects held out for training the TotalCapture+FineTune(H36M Model) model (the additional test pairs were subjects: 1-5, 5-6 6-7 7-8 8-9), with the results presented in Tab: 5. The table illustrates the test on the standard pair of subjects S9 and S11 and the mean and standard deviation from our cross-validation experiment using the model TotalCapture+FineTune(H36M Model). The mean performance across random pairs of test subjects is like that of the official S9/S11 test split, and the variance is low. While the results are not comparable to those in Tab. 4 using the conventional Human3.6M method, they do show the stability of the approach across different unseen test subject pairings.
Table 4 A Comparison of our three approaches compared to other works on the Human 3.6m dataset.

| Approach          | Direct. | Discus | Eat | Greet | Phone | Photo | Pose | Purch. |
|-------------------|---------|--------|-----|-------|-------|-------|------|--------|
| Lin [27]          | 132.7   | 183.6  | 132.4 | 164.3 | 162.1 | 205.9 | 150.6 | 171.3  |
| ekim [44]         | 85.0    | 108.8  | 84.4 | 98.9  | 119.4 | 95.7  | 98.5 | 93.8   |
| Tome [46]         | 65.0    | 73.5   | 76.8 | 86.4  | 86.3  | 110.7 | 68.9  | 74.8   |
| Trumble [50]      | 92.7    | 85.9   | 72.3 | 93.2  | 76.8  | 101.2 | 75.1  | 78.0   |
| Lin [32]          | 58.0    | 68.3   | 63.3 | 65.8  | 75.3  | 93.1  | 61.2  | 65.7   |
| Martinez [31]     | 51.8    | 56.2   | 58.1 | 59.0  | 69.5  | 78.4  | 55.2  | 58.1   |
| Trumble [48]      | 41.7    | 43.2   | 52.9 | 70.0  | 83.9  | 57.3  | 63.5  |        |
| Imtiaz [20]       | 44.2    | 46.7   | 52.3 | 49.3  | 59.9  | 59.4  | 47.5  | 46.2   |
| Human3.6Model     | 55.6    | 52.1   | 51.8 | 59.9  | 62.1  | 58.2  | 55.2  | 62.0   |
| TotalCaptureModel | 37.1    | 45.3   | 47.1 | 45.9  | 60.1  | 57.6  | 49.9  | 48.1   |
| TotalCapture+FineTune (H36M Model) | 36.0 | 44.0 | 43.5 | 43.5 | 53.3 | 58.2 | 47.1 | 45.2 |

Table 5 A Comparison of testing on subjects S9 and S11 (Proposed S9,11) against five-fold cross-validation of other subject pairs on the Human 3.6m dataset.

| Approach          | Direct. | Discus | Eat | Greet | Phone | Photo | Pose | Purch. |
|-------------------|---------|--------|-----|-------|-------|-------|------|--------|
| CrossVal Pairs mean | 40.5   | 49.8   | 43.0 | 53.5  | 51.4  | 56.8  | 50.2 | 49.5   |
| CrossVal Pairs sd  | 3.6    | 5.3    | 4.1 | 7.8   | 3.9   | 4.7   | 6.4  | 5.3    |
| Proposed S9,11     | 36.0   | 44.0   | 43.5 | 43.5  | 53.3  | 58.2  | 47.1 | 42.2   |

5.5 In depth Analysis on TotalCapture dataset

We can explore and analyse additional parameters in the approach in more detail on the TotalCapture dataset. We investigate the effect of increasing the angle between the cameras to vary the camera layout, the input PVH resolution, training data quantity and quality and the effects of the skip connections.

5.5.1 Camera layout

In the two test datasets, we arrange the cameras in a circular wide baseline arrangement of a 360°, however, when using only two cameras, the arrangement of the two cameras will affect the performance of the approach slightly, with neighbouring cameras up to 90° providing the best performance. To demonstrate the effect on the camera placement when C = 2, we create the input PVH from cameras separated by 45°, 90°, 135° and the failure case at 180°. Tab. 6 indicates the average joint error and MSE of the reconstruction for each camera position. The consistent joint and reconstruction results indicate up to 135° indicate the stability of the camera to the camera location and therefore varying quality of input PVHs. At 180° there is very poor performance as the input PVH is very degraded. Fig. 13 also shows the resulting PVH, from increasingly wide baseline cameras, separated by 45°, 90°, 135° and 180°. With 4 cameras used there is not noticeable difference in performance of joint estimator or reconstruction error, making it invariant to camera locations.
5.6 Input PVH Resolution

Our approach uses an input tensor $V_L \in \mathbb{R}^{W \times H \times D \times \phi}$, where $W$, $H$, $D$, and $\phi$ are the width, height, depth and channel of the performance capture volume respectively. For this work, the input volume size is $32^3$, $64^3$ and $128^3$ voxels, the greater the resolution of the volume; the more detail is possible on the reconstruction. This is illustrated in Fig. 14 where the same frame has been reconstruction and the pose estimated for the three input tensor resolutions. Looking at the results, Fig. 14 Qualitative examples of pose estimation and reconstruction using a two-camera input for differing input tensor sizes.

there is no noticeable difference between them, apart from the additional detail for the high-resolution input tensor. The is confirmed with the pose estimation, where the average performance for TotalCapture dataset for $32^3$, $64^3$ and $128^3$ input tensor on the TotalCapture is $21.4\text{mm}$, $21.9\text{mm}$, and $21.5\text{mm}$ respectively, these results are within noise margins. For the results shown in the sections above, we use the $32^3$ tensor, as being the smallest dimensions means fewer parameters to learn and therefore reducing the training time.

5.7 Training Data Quantity and Quality

Generally, for training neural networks a large amount of varied data is required, and the more data the higher the performance, especially as we use 3D convnets, which have an additional dimension and therefore additional filter weights to learn. To learn the weights, the datasets have large amounts of frames available, with the Total-Capture dataset containing over 250,000 frames from each camera view, and Human3.6M’s 500,000 training frames. Therefore, we investigate how the quantity of training data affects the inference performance and also examine the quality or variation of the data. The test sequences were kept consistent throughout as before, and an increasing percentage of total available training data was used from Subjects 1, 2 and 3, randomly sampled from maximum of $\sim 250k$ MVV frames. We also investigate the quality of the data by removing the ROM (range of motion) sequences from the frames we randomly sample over. This is because the ROM sequence is traditionally used by motion capture studios to calibrate their system as it designed to contain a wide range of possible motions. Tab. 7 suggests that the performance is relatively unaffected by the lower amounts of training data providing at least $84\%$ of the joint accuracy with only $20\%$ of the data. Interestingly this can be in part due to the use of our ROM sequences within the training set as when this is not used in the training a further $10-15\%$ performance loss occurs. Therefore, the approach can train using only a sparse set of data and does not over-fit even if only using $20\%$ of the training data.

5.8 Skip connections

The inclusion of skip connections enables details of volume shape to pass through to the final reconstruction in the GAN skipping the bottleneck of the encoder to enable the maintenance of high-frequency information such as the human extremities. Fig. 15 shows an example of the learnt output for a PVH with and without skip connections, the result missing the extremities without the skip connections. There is also a discussion around the process used to combine the skip layer with the rest of the model. ResNet [18], for example, uses element-wise addition to combine the two paths of the model, enabling the model to result in skipping certain layers. It is also possible to average or concatenate the two paths, and Tab. 8 illustrates the performance of our proposed model for both the TotalCapture and Human 3.6M datasets in terms of mean joint error using these three possible processes.

As can be seen from Tab. 8, there is only minor variance between the different methods of combining the skip connection into the main model layers, with mean
and addition, providing similar performance. However, it demonstrates the dramatic error increase with no skip connections as all extremity joints such as arms and legs are incorrectly estimated; we use the mean average for this work.

6 Outdoor footage

To demonstrate the approach in the envisioned scenario of a less constrained capture environment the framework was tested on an internally produced and more challenging dataset, TotalCaptureOutdoor \cite{29}. This is a multi-view video dataset shot outdoors with a moving and varying background of trees and differing illumination. We use 4 of the 6 video cameras placed in a 120° arrangement around the subject, with a large 10x10m capture volume. The subject can be far from the camera and small in the scene as shown in Fig. 16, making traditional 3D pose estimation and volume reconstruction very challenging. There is no ground truth annotation available for TotalCaptureOutdoor and only qualitative results are presented, on two sequences: Subject1, Freestyle and Subject2, Acting1. Using the model trained on the TotalCapture dataset of section 5.1, Fig. 16 illustrates that, despite the small size of the subject in the camera images, an accurate estimation of their pose and volume can be recovered from a coarse initial volume reconstruction. Further demonstrating the robustness of the approach, Fig. 17 illustrates the reconstructed volumes viewed at 60 intervals around the subject. Despite the 4 cameras having only 90° coverage of the scene, the reconstructions are complete and consistent when rendered from unseen viewpoints.

7 Conclusions

This proposed work generates accurate 3D joint and 3D volume proxy reconstructions, from a minimal set of only two wide baseline cameras. Through the learning of a model constrained by a dual loss on the joints and a generative adversarial loss on the 3D volume. The dual loss in conjunction with the Discriminator in the GAN framework delivers state of the art performance. Furthermore, we have demonstrated that a trained model with plentiful data (from the TotalCapture dataset) can be used to improve performance on other sets of data (in this case from the Human3.6M dataset) that have a limited set of camera views.

Acknowledgements

The work was supported by InnovateUK via the Total Capture project, grant agreement 102685 and through the donation of GPU hardware by Nvidia.

References

1. Abrahamsson, S., Blom, H., Jans, D.: Multifocus structured illumination microscopy for fast volumetric super-resolution imaging. Biomedical Optics Express 8(9), 4135–4140 (2017)

| Dataset     | Skip methods, (Mean Joint error) |
|-------------|----------------------------------|
|             | Addition | Average | Concatenation | None     |
| TotalCapture| 55.4mm   | 54.7mm  | 61.2mm       | 98.4mm   |
| H3.6M       | 21.5mm   | 21.4mm  | 24.3mm       | 112.3mm  |

Table 8 A Comparison of skip connection combining for the TotalCapture and Human 3.6M, using the best performing experimental approaches from Tab. 1 and Tab. 4. None indicates no skip connections are used.

Fig. 15 Resulting 3D reconstruction for different skip connections including none for two example frames.

Fig. 16 Accurate Pose and shape reconstruction using challenging distant exterior footage.

Fig. 17 Accurate Pose and shape reconstruction from 7 virtual camera viewpoints in a 360° arrangement around the subject.
Semantic Estimation of 3D Body Shape and Pose using Minimal Cameras

2. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3686–3693 (2014).

3. Aydin, V., Foroosh, H.: Volumetric super-resolution of multispectral data. In: Corr. arXiv:1705.05745v1 (2017).

4. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. ECCV (2016).

5. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR (2017).

6. Chen, C.H., Ramanan, D.: 3d human pose estimation= 2d pose estimation+ matching. In: CVPR (2017).

7. Doug, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Machine Intelligence 38(2), 295–307 (2016).

8. Elhayek, A., de Aguiar, E., Jain, A., Tompson, J., Larlus, D., Morency, L.P., Rehg, J.M., Fua, P.: Learning from multiple views for marker-less motion capture in general scenes with a low number of cameras. In: Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on, pp. 3810–3818 (2015).

9. Fattal, R.: Image upsampling via imposed edge statistics. In: Proc. ACM SIGGRAPH (2007).

10. Gilbert, A., Trumble, M., Malleson, C., Hilton, A., Colombo, J.: Fusing visual and inertial sensors with semantics for 3d human pose estimation. International Journal of Computer Vision 127(4), 381–397 (2019).

11. Gilbert, A., Volino, M., Colombo, J., Hilton, A.: Volumetric performance capture from minimal camera viewpoints. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 556–581 (2018).

12. Gilbert, T.M., C., A.M., A., H., J., C.: Total capture: 3d human pose estimation fusing video and inertial sensors. In: BMVC17 (2017).

13. Glassner, D., Bagon, S., Irani, M.: Super-resolution from a single image. In: Proc. Intl. Conf. Computer Vision (ICCV) (2009).

14. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp. 249–256 (2010).

15. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems, pp. 2672–2680 (2014).

16. Grauman, K., Shakhnarovich, G., Darrell, T.: A bayesian approach to image-based virtual hull reconstruction. In: Proc. CVPR (2003).

17. Guillemaut, J.Y., Hilton, A.: Joint multi-layer segmentation and reconstruction for free-viewpoint video applications. International journal of computer vision 93(1), 73–100 (2011).

18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016).

19. Hochreiter, S., Schmidhuber, J.: Long short-term memory. In: Neural computation, vol. 9, pp. 1735–1780. MIT Press (1997).

20. Hossain, M.R.I., Little, J.J.: Exploiting temporal information for 3d human pose estimation. In: European Conference on Computer Vision, pp. 69–86. Springer (2018).

21. Huang, P., Tjahja, M., Colombo, J., Hilton, A.: Hybrid skeletal-surface motion graphs for character animation from 4d performance capture. ACM Transactions on Graphics (ToG) (2015).

22. 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), 1325–1339 (2014).

23. Jain, V., Seung, H.: Natural image denoising with convolutional networks. In: Proc. Neural Inf. Processing Systems (NIPS), pp. 769–776 (2008).

24. Jiang, H.: Human pose estimation using consistent max-covering. In: Intl. Conf. on Computer Vision (2009).

25. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Proc. NIPS (2012).

26. Lan, X., Hutenlocher, D.: Beyond trees: common-factor model for 2d human pose recovery. In: Proc. Intl. Conf. on Computer Vision, vol. 1, pp. 470–477 (2005).

27. Li, S., Zhang, W., Chan, A.B.: Maximum-margin structured learning with deep networks for 3d human pose estimation. In: The IEEE International Conference on Computer Vision, pp. 2848–2856 (2015).

28. 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), 248 (2015).

29. Malleson, C., Gilbert, A., Trumble, M., Colombo, J., Hilton, A.: Real-time full-body motion capture from video and imus. In: 3DV (2017).

30. von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 601–617 (2018).

31. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. ICCV (2017). URL https://arxiv.org/abs/1705.03098

32. Mude Lin Liang Lin, X.L.K.W., Cheng, H.: Recurrent 3d pose sequence machines. In: CVPR (2017).

33. Park, D., Ramanan, D.: Articulated pose estimation with tiny synthetic videos. In: Proc. CHA-LEARN Workshop on Looking at People (2015).

34. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3d human pose. In: CVPR (2017).

35. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Harvesting multiple views for marker-less 3d human pose annotations. In: CVPR (2017).

36. Ren, R., Colombo, J.: Visual sentences for pose retrieval over low-resolution cross-media dance collections. IEEE Transactions on Multimedia (2012).

37. Ren, X., Berg, E., Malik, J.: Recovering human body configurations using pairwise constraints between parts. In: Proc. Intl. Conf. on Computer Vision, vol. 1, pp. 824–831 (2005).

38. Rhodin, H., Robertini, N., Cañas, D., Richardsd, C., Seidel, H.P., Theobalt, C.: General automatic human shape and motion capture using volumetric contour cues. In: European Conference on Computer Vision, pp. 509–526. Springer (2016).

39. Rudin, L.I., Osher, S., Fatemi, E.: Non-linear total variation based noise removal algorithms. Physics D 60(1-4), 259–268 (1992).

40. Sanzari, M., Tsouklos, V., Pirri, F.: Bayesian image based 3d pose estimation. In: European Conference on Computer Vision, pp. 566–582. Springer (2016).

41. 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 sub-pixel convolutional neural network. In: Proc. Comp. Vision and Pattern Recognition (CVPR) (2016).

42. Srivastava, P., Shi, J.: Bottom-up recognition and parsing of the human body. In: Proc. Computer Vision and Pattern Recognition, pp. 1–8 (2007).

43. Srivastava, R.K., Greff, K., Schmidhuber, J.: Training very deep networks. In: Advances in neural information processing systems, pp. 2377–2385 (2015).
44. 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)
45. Tekin, B., Rozantsev, A., Lepetit, V., Fua, P.: Direct prediction of 3d body poses from motion compensated sequences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 991–1000 (2016)
46. Tome, D., Russell, C., Agapito, L.: Lifting from the deep: Convolutional 3d pose estimation from a single image. arXiv preprint arXiv:1701.00295 (2017)
47. Toshev, A., Szegedy, C.: Deep pose: Human pose estimation via deep neural networks. In: Proc. CVPR (2014)
48. Trumble, M., Gilbert, A., Hilton, A., Collomosse, J.: Deep autoencoder for combined human pose estimation and body model upscaling. In: European Conference on Computer Vision (ECCV’18) (2018)
49. Trumble, M., Gilbert, A., Hilton, A., John, C.: 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)
50. 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, pp. 1–13. URL http://epubs.surrey.ac.uk/841740/
51. 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), pp. 370–378 (2015)
52. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: CVPR (2016)
53. Xie, J., Xu, L., Chen, E.: Image denoising and inpainting with deep neural networks. In: Proc. Neural Inf. Processing Systems (NIPS), pp. 350–358 (2012)
54. Zeiler, M.D.: Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701 (2012)
55. 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, pp. 4966–4975 (2016)
56. Zhu, Y., Zhang, Y., Yuille, A.L.: Single image super-resolution using deformable patches. In: Proc. Comp. Vision and Pattern Recognition (CVPR), pp. 2917–2924 (2014)