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

Andrew Gilbert[1]
https://www.surrey.ac.uk/people/andrew-gilbert
Matt Trumble[1]

Adrian Hilton[1]
https://www.surrey.ac.uk/people/adrian-hilton
John Collomosse[12]
http://personal.ee.surrey.ac.uk/Personal/J.Collomosse/index.php

[1] Centre for Vision Speech and Signal Processing,
University of Surrey,
Guildford,
UK

[2] Creative Intelligence Lab,
Adobe Research, USA

Abstract

We aim to simultaneously estimate the 3D articulated pose and high fidelity volumetric occupancy of human performance, from multiple viewpoint video (MVV) with as few as two views. We use a multi-channel symmetric 3D convolutional encoder-decoder with a dual loss to enforce the learning of a latent embedding that enables inference of skeletal joint positions and a volumetric reconstruction of the performance. The inference is regularised via a prior learned over a dataset of view-ablated multi-view video footage of a wide range of subjects and actions, and show this to generalise well across unseen subjects and actions. We demonstrate improved reconstruction accuracy and lower pose estimation error relative to prior work on two MVV performance capture datasets: Human 3.6M and TotalCapture.

Only 2 camera views

Accurate 3D pose and shape
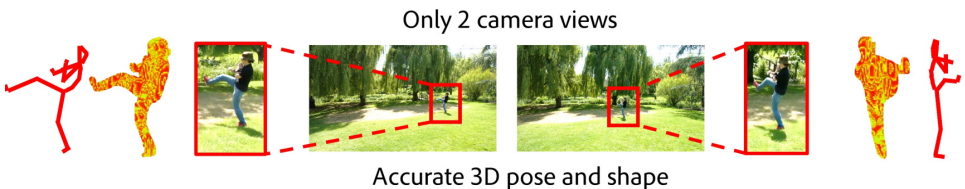
1 Introduction

Human performance capture is used extensively within biomechanics and the creative industries. Commercial approaches are typically constrained to skeletal joint estimation in the presence of subject-worn markers captured from multiple viewpoints by specialised (e.g. infra-red) cameras. In this paper, we present a method for video-based performance capture, able to estimate both 3D skeletal pose and shape (volumetric occupancy) of a subject accurately from multiple-viewpoint video (MVV). Uniquely, we do so without a parametric shape model (*e.g.* SMPL [21]), and without the need for worn markers or sensors [36, 42], nor a large camera count [8]. Our approach considers MVV with as few as two wide baseline cameras, motivated by real-world scenarios that may constrain the on-set deployment of

large numbers of witness camera views due to limitations on camera cost or placement (*e.g.* security or sports events).

Our technical contribution 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. In a single inference step, we estimate both the skeletal joint positions (pose) and refine a higher fidelity volumetric reconstruction from the rough proxy (occupancy).

Our architecture is a volumetric encoder-decoder convolutional neural network (CNN) in which 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. A joint loss between both outputs is used within a generative adversarial network to ensure 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. Unlike prior work that has explored volumetric encoder-decoder networks for pose [38] or for content up-scaling [12], we leverage use of 2D semantic detections to supplement the background occupancy volumetric proxy. The encoder-decoder network serves to learn a prior for human shape, regularised by a generative adversarial network (GAN) loss that ensures realism in the output high-fidelity volumetric reconstruction output and enabling both the pose estimation and reconstruction to be learnt from a minimal set of camera views. The work by Trumble *et al* [38] inspires this work, however with significantly improved performance through the introduction of several notable novelties; the inclusion of semantic labels as well as occupancy probabilities in the voxels that make the PVH. The incorporation of a GAN discriminator on the output volume and the extension of the bottleneck of the encoder-decoder with additional latent features besides the body joint coordinates. We demonstrate SOTA results and several ablation studies in the paper which show the value of these contributions.

2   Related Work

Our work is inspired by contemporary super-resolution (SR) algorithms that apply learned priors to enhance visual detail in images, volumetric performance capture or reconstruction and human pose estimation (HPE).

Super-resolution: Classical image restoration / SR approaches combine multiple data sources (*e.g.* images [11], or self-similar patches [13, 47]) under regularization *e.g.* total variation [29]. Convolutional neural network (CNN) autoencoders have been applied to image [9, 43, 45] and video-upscaling [30]. Volumetric SR has been explored for microscopy [6], and for multi-spectral sensing [4]. Recently SR for volumetric performance capture was explored using encoder-decoder networks [12].

Volumetric Performance Reconstruction: Volumetric performance capture pipelines typically use multiple wide baseline viewpoints [7, 32] arranged around the capture volume. More recently, data driven machine learning approaches [39, 40, 46] have demonstrated improve reconstruction from a single camera. Varol *et al* [40] use a neural network for direct inference of volumetric body shape from a single image. While Jackson *et al* [39] directly regress the volumetric representation of the 3D geometry using a standard, spatial, CNN architecture, and Zheng *et al* [46] also uses the parametric representation of the SMPL body model [21] fusing different scales of image features into the 3D space through volumetric feature transformation, to recover accurate surface geometry.

Human Performance Estimation: There are two distinct categories of HPE; bottom-up data-driven and top-down fitting a model. In general, top-down 2D pose estimation fits 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
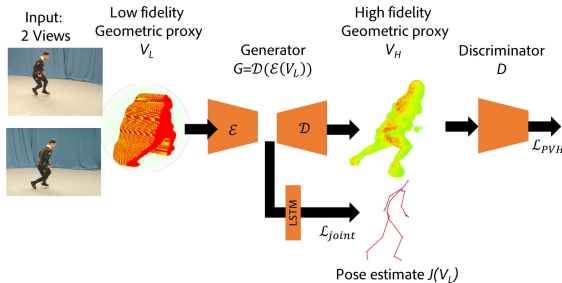
Figure 1: Network architecture. The input is a low fidelity geometric proxy (V_L) from two wide baseline camera views. This proxy is passed through a decoder-encoder ($D, \mathcal{E}$) to produce a 3D human pose estimate (joint angles $J(V_L)$) via the latent space and to output, a high-fidelity geometric proxy (V_H) regularised via discriminator (D).

data defined model such as the SMPL Body Model [21]. Spatio-temporal tracking of pictorial structures is applied to HPE in [19], and [2] explored the fusion of pictorial structures with Ada-Boost shape classification. Malleson *et al* [22] used IMUs with a full kinematic solve to adequately estimate 3D pose both indoor and outdoor. Recently, the SMPL model has been employed by several pose estimation techniques with IMUs [41, 42] and 2D images [17, 53].

Bottom-up pose estimation is driven by image parsing to isolate components, Srinivasan *et al* [51] used graph-cuts to parse a subset of salient shapes from an image and group these into a model of a person. Ren *et al* [27] recursively splits Canny edge contours into segments, classifying each as a putative body part using cues such as parallelism. Ren [26] also used Bag of Visual Words for implicit pose estimation as part of a pose similarity system for dance video retrieval. In DeepPose, Toshev [55] used a cascade of convolutional neural networks to estimate 2D pose in images. Elhayek *et al* [10] used MVV with a Convnet to produce 2D pose estimations while Rhodin *et al* [28] minimised the edge energy inspired by volume ray casting to deduce the 3D pose. Trumble *et al* [57] used a flattened MVV based spherical histogram with a 2D convnet to estimate pose. While Pavlakos *et al* [24] used a simple volumetric representation in a 3D convnet for pose estimation and Wei *et al* [44] performed related work in aligning pairs of joints to estimate 3D human pose. Since detecting pose for each frame individually leads to incoherent and jittery predictions over a sequence, many approaches exploit temporal information [3, 23] often using LSTMs [15]. Trumble et al. [58] estimate 3D pose using the latent space of a volumetric encoder-decoder, but do not incorporate semantic information nor GAN constraint.

3 Joint minimal camera Pose and Volume reconstruction

We present an overview of our process for simultaneously estimating pose and high fidelity occupancy in Figure 1. First, a pre-processing step [14] reconstructs a coarse Probabilistic Visual Hull (PVH) proxy using a limited number of cameras (Sec. 3.2). For each voxel, we encode a feature reflecting its occupancy and semantic label (e.g. joints) lifted from 2D. This initial estimate (Sec. 3.1) typically contains phantom limbs and sub-volumes. Next, a 3D convolutional encoder-decoder (Sec. 3.3) and generative adversarial network (GAN) (Sec. 3.3.1), 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, while the 3D joint estimations are fed to LSTM layers to enforce the temporal consistency of the 3D joints (Sec. 3.3.3).

3.1 Visual Features

To estimate the pose, we propose to lift 2D visual features to form a 3D voxel features from two distinct modes created from RGB images of each camera view; 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 I and a predefined clean plate P approximates pixel occupancy. A thresholded $L2$ distance between the two images in the HSV colour domain provides a soft occupancy probability. 2D joint belief labels estimated through the approach of Wei [6, 44] generate the 2D semantic joint detections, a multi-stage process that iteratively refines the 2D joint estimates based on both the input image and the previous stage's returned pixel-wise belief map. At each stage s and for each joint label j the algorithm returns dense per pixel belief maps m_s^j, which provides the confidence of a joint centre for any given pixel (x, y).

$$M(x,y) = \arg\max_j m_s^j(x,y) \tag{1}$$

The per joint belief maps are maximised over the confidence of all possible joint labels to produce a single label per pixel image $M(x,y)$.

3.2 Volumetric Representation

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 $c = [1, .., C]$ 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 subvolume $X_I \in \mathcal{V}$ corresponding to, and centred upon, a single subject, and which is decimated into voxels $\mathbf{V_L}^i = [\begin{array}{ccc} v_x^i & v_y^i & v_z^i \end{array}]$ for $i = [1, \ldots, |\mathbf{V_L}|]$; each voxel is 5mm^3 in size. Each voxel $v^i \in \mathbf{V_L}$ projects to coordinates $(x[v^i], y[v^i])$ in each camera view c.

Then given an 2D image denoted as I_c, with $\Phi = [1, \ldots, \phi]$ feature channels (from 2D occupancy/joints), point (x_c, y_c) is the point within I_c to which $\mathbf{V_L}^i$ projects in a given view:

$$x[\mathbf{V_L}^i] = \frac{f_c v_x^i}{v_z^i} + o_c^x \quad \text{and} \quad y[\mathbf{V_L}^i] = \frac{f_c v_y^i}{v_z^i} + o_c^y, \tag{2}$$

$$[\begin{array}{ccc} v_x^i & v_y^i & v_z^i \end{array}] = COP_c - R_c^{-1} V_L^i. \tag{3}$$

The likelihood of the voxel being part of the performer in a given view c is:

$$p(\mathbf{V_L}^i|c) = I_c(x[\mathbf{V_L}^i], y[\mathbf{V_L}^i], \phi). \tag{4}$$

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

$$p(\mathbf{V_L}^i, \phi) = \prod_{i=1}^{C} 1/(1 + e^{-p(\mathbf{V_L}^i|c)}). \tag{5}$$

3.3 Dual Loss Convolutional Volumetric Network

We propose to learn a deep representation or output given an input tensor $\mathbf{V_L}$ where $\mathbf{V_L} \in \mathbb{R}^{X \times Y \times Z \times \phi}$, where each dimension encodes the probability of volume occupancy $p(X,Y,Z)$ derived from a PVH obtained using a low camera count (Eq.5) from channels (ϕ); foreground occupancy and semantic 2D joint estimates. We wish to train a deep representation to solve the prediction problem $\mathbf{V_H} = \mathcal{F}(\mathbf{V_L})$ for similarly encoded tensor $\mathbf{V_H} \in \mathbb{R}^{W \times H \times D \times \phi}$ derived from a higher fidelity PVH of identical dimension obtained using a higher camera count. Where W,H,D,ϕ are the width, height, depth and channel of the performance capture volume respectively. Function $\mathcal{F}$ 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 $\mathbf{V_H}$, where $\mathbf{V_H} = \mathcal{F}(\mathbf{V_L}) = \mathcal{D}(\mathcal{E}(\mathbf{V_L}))$ for the learnt encoder ($\mathcal{E}$) and decoder ($\mathcal{D}$) 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 $J(\mathbf{V_L}) = \mathcal{E}(\mathbf{V_L})$ where $J(\mathbf{V_L})$ 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 $\mathbf{e}$ (in general $\mathbf{e} = 200$). The decoder inverts this process to output tensor $\mathbf{V_H}$ matching the input resolution but with higher fidelity. The full network parameters are: $n_{\mathcal{E}} = [64,64,128,128,256]$, $n_{\mathcal{D}} = [256,128,128,64,64]$, $k_{\mathcal{E}} = [3,3,3,3,3]$, $k_{\mathcal{D}} = [3,3,3,3,3]$, $k_s = [0,1,0,1,0]$ where $k[i]$ indicates the kernel size and $n[i]$ is the number of filters at layer i for the encoder ($\mathcal{E}$) and decoder ($\mathcal{D}$) parameters respectively. The location of the two skip connections are indicated by s and link two groups of convolutional layers to their corresponding mirrored up-convolutional layer. The passed convolutional feature maps are averaged to the up-convolutional feature maps element-wise and passed to the next layer after rectification.

The goal of $\mathcal{F}$ is thus to regress a high fidelity 3D volumetric representation given an initial poor fidelity blocky 3D volume estimate. Learning the end-to-end mapping from blocky volumes generated from a small number of camera viewpoints to both cleaner high fidelity volumes as if made by a greater number of camera viewpoints and accurate 3D joint position estimates, requires estimation of the weights ϕ in $\mathcal{F}$ represented by the convolutional and deconvolutional kernels. Specifically, given a collection of training sample triplets x^i, z^i, j^i, where $x^i \in \mathbf{V_L}$ is an instance of a low camera count volume, $z^i \in \mathbf{V_H}$ is the high camera count output groundtruth volume and $j^i \in \mathbf{J}$ is a vector of groundtruth joint positions for the given volume. The Mean Squared Error (MSE) is minimised at the output of the bottleneck and decoder across $N = W \times H \times D$ voxels through the two losses $\mathcal{L}_{joint}$ and $\mathcal{L}_{PVH}$.

$$\mathcal{L}(\phi) = \mathcal{L}_{joint} + \lambda \mathcal{L}_{PVH} = \frac{1}{N} \sum_{i=1}^{N} \|\mathcal{F}(x^i : \phi) - z^i\|_2^2 + \lambda \mathcal{E}(\mathbf{V_L} : \phi) - j^i\|_2^2 \quad (6)$$

Where $\lambda = 10^-3$, ensures both terms are of a similar magnitude.

3.3.1 Generative Adversarial Network Model

The encoder-decoder model described in the section above with the dual volume and joint pose loss can produce consistent 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 (GAN).

The encoder model from section 3.3, which we refer to as the *Generator G* estimates the improved volume, whilst the discriminator maximises the chance of recognising real PVH

volumes as real and generated PVH volumes as fake, optimizing the minimax objective::

$$\min_{G} \max_{D} V(D,G) = \mathbb{E}_{x \sim \mathbb{P}_r}[\log(D(x))] + \mathbb{E}_{\tilde{x} \sim \mathbb{P}_g}[\log(1 - D(\tilde{x}))] \tag{7}$$

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.

3.3.2 Skip Connections

Deeper networks in image restoration tasks can result in finer image details being lost given the compact latent space. Recovery of this detail is an under-determined problem, exasperated by the need to reconstruct the additional dimension in volumetric data. We add skip connections between two corresponding convolutional and deconvolutional layers. Omitting the skip connections the detail of extremities such as lower arm position is poorly estimated by both the volume and 3d joints (see sup. material).

3.3.3 Temporal Consistency

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

4 Results and Discussion

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. We performed quantitative evaluation over two public multi-view video datasets of human actions. 3D human pose is evaluated for Human 3.6M [13], and the performance of both the skeleton estimation and volume reconstruction is evaluated on TotalCapture [56].

To train $\mathcal{F}$, we initially, train the encoder for just the skeleton loss, purely as a pose regression task without the decoder or critic networks, due to the large parameter count in the volumetric network. 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.

We train with a batch size of 32 and a sequence length of $T = 5$ (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 central vertical axis of the PVH to introduce rotation invariance.

Approach	Num Cams	SeenSubjects(S1,2,3)			UnseenSubjects(S4,5)			Mean
		W2	FS3	A3	W2	FS3	A3	
Tri-CPM-LSTM [5]	8	45.7	102.8	71.9	57.8	142.9	59.6	80.1
2D Matte-LSTM [57]	8	94.1	128.9	105.3	109.1	168.5	120.6	121.1
3D-PVH [56]	8+13 IMU	30.0	90.6	49.0	36.0	112.1	109.2	70.0
AutoEnc [58]	8	13.4	49.8	24.3	22.0	71.7	40.7	35.5
Fusion-RPSM [25]	8	19	58	21	32	54	33	29
IMU 1Cam SMPL [42]	1+13 IMU	-	-	-	-	-	-	26.0
Proposed DualLoss GAN	2	9.2	30.3	15.2	13.3	41.7	25.3	21.4

Table 1: Comparison of our approach on TotalCapture to other human pose estimation approaches, expressed as average per joint error (mm) on previously seen and unseen test subjects. (where W2, FS3, A3 are groups of test sequences of walking, freestyle and acting respectively)

4.1 TotalCapture Evaluation

We quantitatively evaluate tracking accuracy on the TotalCapture dataset [56]. We study the accuracy gain due to our method by ablating the set of camera views available on the *TotalCapture* dataset. Jointly training the generative adversarial dual loss model using high fidelity PVHs obtained using all ($C = 8$) views of the dataset and 78-D vector concatenation of the 26 3D Cartesian pose joint coordinates. With the corresponding input low fidelity, PVHs obtained using fewer views (we train for $C = 2$ and $C = 4$ random neighbouring views), we follow the train and test strategy of [56]. The dataset contains five subjects, with four diverse categories of sequences; *ROM, Walking, Acting, and Freestyle*, with each sequence, repeated three times by each subject. The sequences are long, with around 3000-5000 frames, resulting in 1.9M frames. Within the acting and freestyle sequences, there is a great deal of diversity in the captured content.

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. 4). 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 unabated $C = 8$ camera viewpoints. Tab. 1 quantifies the pose animation error between previous approaches using in general multiple camera views [5, 56, 57, 58] or additional data modalities [56, 42] and our proposed approach with only two camera views. We outperform best camera approach [25] by 8 mm indicating the importance of the GAN loss and semantic 2D joint estimates.
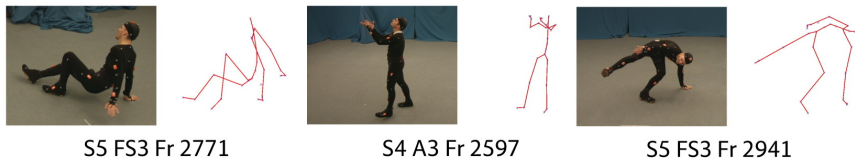

S5 FS3 Fr 2771   S4 A3 Fr 2597   S5 FS3 Fr 2941

Figure 2: Representative pose estimations from (Fr)ames of unseen (S)ubjects performing (A)ctions with challenging poses (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	-	✓	-	-	-	42.0	120.5	59.8	58.4	162.1	103.4	85.4
EncoderLSTM	8cam	-	✓	-	✓	-	15.2	65.7	54.4	17.8	73.0	50.6	58.4
AutoEncLSTM	8cam	-	✓	✓	✓	-	13.4	49.8	24.3	22.0	71.7	40.7	35.5
2DJoint	-	8cam	✓	✓	✓	-	21.2	123.1	88.6	105.7	142.2	97.7	41.2
Occ+2DJoint	8cam	8cam	✓	✓	✓	-	10.2	123.1	88.6	105.7	142.2	97.7	31.1
GAN8cam	8cam	8cam	✓	✓	✓	✓	8.2	30.5	15.0	10.2	40.8	24.7	20.7
GAN4cam	4cam	4cam	✓	✓	✓	✓	9.8	29.9	15.3	13.5	42.2	24.9	21.6
GAN2cam	2cam	2cam	✓	✓	✓	✓	9.2	30.3	15.2	13.3	41.7	25.3	21.4

Table 2: Ablation study of the Mean per joint error (mm). for the individual components on the TotalCapture dataset.
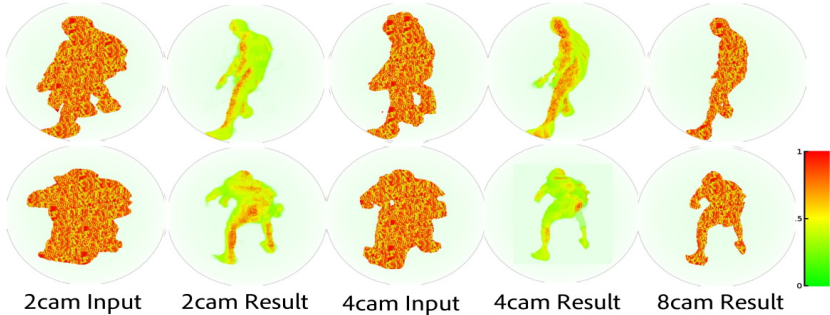

Figure 3: Examples of input/resultant reconstructions for [2,4,8] cameras on TotalCapture.

4.2 Ablation Study

To understand the influence of the individual components and design decisions, we perform an ablative analysis of tracking accuracy for our individual contributions (Tab. 2). Each part of the process enables an improvement in the accuracy performance, especially the use of temporal information (**EncoderLSTM**) and dual loss in the approach (**AutoEncLSTM**). 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 eight camera views. The greater the number of cameras, the more visually realistic the input PVH is. However, it is possible to remove a large number of these cameras with little or no impact on performance (**GAN4cam** and **GAN2cam**). Despite greatly degrading the appearance of the input PVH when using only 2 or 4 views as input, as indicated by Fig. 3. 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, using a low fidelity PVH from only two camera views with phantom and missing voxels, achieves a headline performance of 21.4mm mean per joint error.

4.3 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

| Method | Cams | SeenSubs(S1,2,3) | | | UnseenSubs(S4,5) | | | Mean |
	C	W2	FS3	A3	W2	FS3	A3	
Input	2	19.1	28.5	23.9	23.4	27.5	25.2	24.6
Input	4	11.4	16.5	12.5	12.0	15.2	14.2	11.6
[12]	2	5.43	10.03	6.70	5.34	10.05	8.71	7.71
Ours	2	5.44	9.94	6.34	5.16	9.86	8.49	7.34
Ours	4	4.85	9.32	5.84	4.83	9.56	8.03	7.02

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 $\times 10^{-3}$). Our method reduces reconstruction error to 30% of the baseline (Input) for two views.

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. View count $C = 4$ 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. 4.
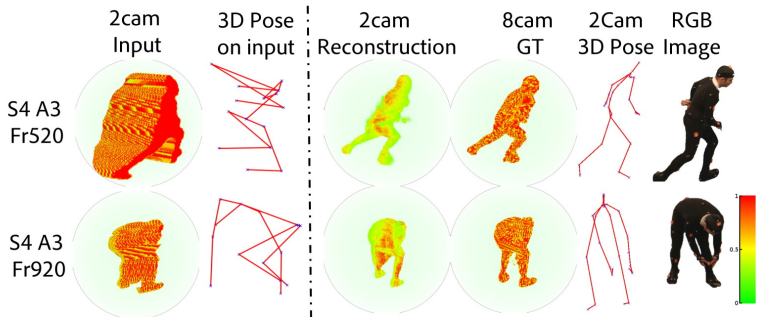

Figure 4: Qualitative visual comparison of the input PVH and 3D Pose estimate on encoder against the resultant Reconstruction and 3D Pose estimation using $C = \{2\}$ views on the TotalCapture dataset. False colour volume occupancy (PVH) and groundtruth $C = 8$ PVH.

4.4 Human 3.6M evaluation

We perform a further quantitative and qualitative evaluation on the Human 3.6M [18] 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. Each video is captured using four calibrated cameras arranged in the $360°$ arrangement and contains 3D pose ground truth. We follow the standard train and evaluation protocols of the Human3.6M dataset [20, 34]. Therefore, we explore (Tab. 4) 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:

Approach	Direct.	Discus	Eat	Greet.	Phone	Photo	Pose	Purch.
Lin *et al* [20]	132.7	183.6	132.4	164.4	162.1	205.9	150.6	171.3
Lin *et al* [23]	58.0	68.3	63.3	65.8	75.3	93.1	61.2	65.7
Trumble *et al* [58]	41.7	43.2	52.9	70.0	64.9	83.0	57.3	63.5
Imtiaz *et al* [16]	44.2	46.7	52.3	49.3	59.9	59.4	47.5	46.2
Qiu *et al* [25]	28.9	32.5	26.6	28.1	28.3	29.3	28.0	36.8
Human3.6Model	55.6	52.1	51.8	59.9	62.1	58.2	55.2	62.0
TCModel	37.1	45.3	47.1	45.9	60.1	57.6	49.9	48.1
TCModel+FineTune(H36M)	26.0	24.0	23.5	23.5	33.3	38.2	27.1	25.2
	Sit.	Sit D	Smke	Wait	W.Dog	walk	W. toget.	Mean
Lin *et al* [20]	151.6	243.0	162.1	170.7	177.1	96.6	127.9	162.1
Lin *et al* [23]	98.7	127.7	70.4	68.2	73.0	50.6	57.7	73.1
Trumble *et al* [58]	61.0	95.0	70.0	62.3	66.2	53.7	52.4	62.5
Imtiaz *et al* [16]	59.9	65.6	55.8	50.4	52.3	43.5	45.1	51.9
Qiu *et al* [25]	42.0	30.5	35.6	30.0	28.3	30.0	30.5	31.2
Human3.6Model	53.3	74.6	61.8	59.1	61.8	65.8	61.2	59.6
TCModel	56.8	68.2	56.3	53.1	47.7	50.5	50.2	54.7
TCModel+FineTune(H36M)	30.2	48.1	37.6	31.2	34.4	28.1	27.1	30.5

Table 4: Comparison of the proposed 3 methods to baseline methods on Human 3.6M.

Human3.6Model: A baseline approach, using the specified Human 3.6M training data with the four cameras assuming the semantic 2D joints will compensate in part for the phantom part and ghosting that occurs to the occupancy voxels.
TCModel: Transfer of the trained $2 \mapsto 8$ camera views model from the TotalCapture dataset, without any further training, to estimate pose as if 8 cameras were used at acquisition.
TCModel+FineTune(H36M): 2 epochs of fine-tuning of the learnt $2 \mapsto 8$ TCModel on Human3.6M dataset.
Our 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)** improves this performance by a further 10mm. Our network improves on Qiu [25], and dramatically improves on other prior work. 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 for previous methods did not perform well due to heavy occlusion.

5   Conclusions

This proposed work generates accurate 3D joint and 3D volume proxy reconstructions, from a minimal set of only two wide baseline cameras, through learning 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.

References

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

4140, 2017.

[2] M. Andriluka, S. Roth, and B. Schiele. Pictoral structures revisited: People detection and articulated pose estimation. In *Proc. Computer Vision and Pattern Recognition*, 2009.

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[4] V. Aydin and H. Foroosh. Volumetric super-resolution of multispectral data. In *Corr. arXiv:1705.05745v1*, 2017.

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *ECCV'16*, 2016.

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[7] Dan Casas, Peng Huang, and Adrian Hilton. Surface-based Character Animation. In Marcus Magnor, Oliver Grau, Olga Sorkine-Hornung, and Christian Theobalt, editors, *Digital Representations of the Real World: How to Capture, Model, and Render Visual Reality*, chapter 16, pages 239–252. CRC Press, April 2015. ISBN 9781482243819.

[8] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (TOG)*, 34(4):69, 2015.

[9] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Machine Intelligence*, 38(2):295–307, 2016.

[10] Ahmed Elhayek, Edilson de Aguiar, Arjun Jain, Jonathan Tompson, Leonid Pishchulin, Micha Andriluka, Chris Bregler, Bernt Schiele, and Christian Theobalt. Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3810–3818, 2015.

[11] R. Fattal. Image upsampling via imposed edge statistics. In *Proc. ACM SIGGRAPH*, 2007.

[12] Andrew Gilbert, Marco Volino, John Collomosse, and Adrian Hilton. Volumetric performance capture from minimal camera viewpoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 566–581, 2018.

[13] D. Glasner, S. Bagon, and M. Irani. Super-resolution from a single image. In *Proc. Intl. Conf. Computer Vision (ICCV)*, 2009.

[14] K. Grauman, G. Shakhnarovich, and T. Darrell. A bayesian approach to image-based visual hull reconstruction. In *Proc. CVPR*, 2003.

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. In *Neural computation*, volume 9, pages 1735–1780. MIT Press, 1997.

[16] Mir Rayat Imtiaz Hossain and James J Little. Exploiting temporal information for 3d human pose estimation. In *European Conference on Computer Vision*, pages 69–86. Springer, 2018.

[17] Yinghao Huang, Federica Bogo, Christoph Classner, Angjoo Kanazawa, Peter V. Gehler, Ijaz Akhter, and J. Black. Towards accurate markerless human shape and pose estimation over time. In *3DV*, 2017.

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

[19] X. Lan and D. Huttenlocher. A unified spatio-temporal articulated model for tracking. In *Proc. Computer Vision and Pattern Recognition*, volume 1, pages 722–729, 2004.

[20] Sijin Li, Weichen Zhang, and Antoni B Chan. Maximum-margin structured learning with deep networks for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2848–2856, 2015.

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

[22] C Malleson, A Gilbert, M Trumble, J Collomosse, and A Hilton. Real-time full-body motion capture from video and imus. In *3DV*, 2017.

[23] Xiaodan Liang Keze Wang Mude Lin, Liang Lin and Hui Cheng. Recurrent 3d pose sequence machines. In *CVPR*, 2017.

[24] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *CVPR*, 2017.

[25] Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. Cross view fusion for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[26] R Ren and J Collomosse. Visual sentences for pose retrieval over low-resolution cross-media dance collections. *IEEE Transactions on Multimedia*, 2012.

[27] X. Ren, E. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. In *Proc. Intl. Conf. on Computer Vision*, volume 1, pages 824–831, 2005.

[28] Helge Rhodin, Nadia Robertini, Dan Casas, Christian Richardt, Hans-Peter Seidel, and Christian Theobalt. General automatic human shape and motion capture using volumetric contour cues. In *European Conference on Computer Vision*, pages 509–526. Springer, 2016.

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

[30] W. Shi, J. Caballero, F. Huszar, J. Totz, A. Aitken, R. Bishop, D. Rueckert, and Z. Wang. 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.

[31] P. Srinivasan and J. Shi. Bottom-up recognition and parsing of the human body. In *Proc. Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[32] Jonathan Starck, Joe Kilner, and Adrian Hilton. A free-viewpoint video renderer. *Journal of Graphics, GPU, and Game Tools*, 14(3):57–72, 2009.

[33] J Tan, I Budvytis, and R Cipolla. Indirect deep structured learning for 3d human body shape and pose prediction. In *BMVC*, 2017.

[34] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. *arXiv preprint arXiv:1701.00295*, 2017.

[35] A. Toshev and C. Szegedy. Deep pose: Human pose estimation via deep neural networks. In *Proc. CVPR*, 2014.

[36] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *Proceedings of 28th British Machine Vision Conference*, pages 1–13. URL http://epubs.surrey.ac.uk/841740/.

[37] Matthew Trumble, Andrew Gilbert, Adrian Hilton, and Collomosse John. 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.

[38] Matthew Trumble, Andrew Gilbert, Adrian Hilton, and John Collomosse. Deep autoencoder for combined human pose estimation and body model upscaling. In *European Conference on Computer Vision (ECCV'18)*, 2018.

[39] Gül Varol, Duygu Ceylan, Bryan C. Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *In ECCV'18*, 2018.

[40] Gül Varol, Duygu Ceylan, Bryan C. Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *In ECCV'18*, 2018.

[41] Timo von Marcard, Bodo Rosenhahn, Michael Black, and Gerard Pons-Moll. 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.

[42] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018.

[43] Z. Wang, D. Liu, J. Yang, W. Han, and T. S. Huang. Deep networks for image super-resolution with sparse prior. In *Proc. Intl. Conf. Computer Vision (ICCV)*, pages 370–378, 2015.

[44] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016.

[45] J. Xie, L. Xu, and E. Chen. Image denoising and inpainting with deep neural networks. In *Proc. Neural Inf. Processing Systems (NIPS)*, pages 350–358, 2012.

[46] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *In ICCV'19*, 2019.

[47] Y. Zhu, Y. Zhang, and A. L. Yuille. Single image super-resolution using deformable patches. In *Proc. Comp. Vision and Pattern Recognition (CVPR)*, pages 2917–2924, 2014.