

# DeepFuse: An IMU-Aware Network for Real-Time 3D Human Pose Estimation from Multi-View Image

Fuyang Huang

Ailing Zeng

Minhao Liu

Qiuxia Lai

Qiang Xu

The Chinese University of Hong Kong

fyhuang, alzeng, mhliu, qxla, qxu@cse.cuhk.edu.hk

## Abstract

In this paper, we propose a two-stage fully 3D network, namely **DeepFuse**, to estimate human pose in 3D space by fusing body-worn Inertial Measurement Unit (IMU) data and multi-view images deeply. The first stage is designed for pure vision estimation. To preserve data primitiveness of multi-view inputs, the vision stage uses multi-channel volume as data representation and 3D soft-argmax as activation layer. The second one is the IMU refinement stage which introduces an IMU-bone layer to fuse the IMU and vision data earlier at data level. without requiring a given skeleton model *a priori*, we can achieve a mean joint error of 28.9mm on TotalCapture dataset and 13.4mm on Human3.6M dataset under protocol 1, improving the SOTA result by a large margin. Finally, we discuss the effectiveness of a fully 3D network for 3D pose estimation experimentally which may benefit future research.

## 1. Introduction

As a fundamental technique for many applications (e.g., Virtual Reality (VR), Human-Computer Interaction (HCI), and animation making), human pose estimation is a long-standing research problem and has received significant attention from both academia and industry.

While 2D human pose estimation has been extensively studied in the literature (thanks to the availability of large manually-annotated datasets), existing 3D human pose estimation techniques still have many limitations. Marker-based vision solutions (e.g., Vicon [49]) are able to achieve high accuracy in recovering 3D human pose and position, but they require sophisticated setup for the surrounding cameras as well as carefully-calibrated markers on human body. Markerless vision solutions (e.g., Kinect and Leap-Motion) are handier, but they can only capture human pose within a near range and fail when there is occlusion. Alternatively, body-worn IMUs (e.g., Xsens [56]) show remarkable stability and accuracy in capturing bone orientation, but they cannot tell the accurate joint positions.

Considering the pros and cons of the two types of sen-

sors, an interesting problem is whether we could fuse the IMU data and vision data to achieve better results. One challenging issue for such fusion is that the vision input is in pixel/voxel format while the IMU input is in quaternion form. The difference in feature spaces makes it difficult to directly concatenate them into a single network. Trumble *et al.* [47] simply fuse the results from the two kinds of sensors with a fully connected layer at the end of the network for regression. Such a straightforward solution does not realize the potential benefits of combining the two modalities. To tackle this problem, some optimization-based solutions try to fuse the data by introducing pre-defined skeleton lengths [36, 26, 51]. These solutions, however, requires pre-defined skeleton data, and thus they cannot be well generalized for unknown subjects.

To overcome the limitations of the above fusion solutions, we propose **DeepFuse**, a novel IMU-aware network that can fuse the two modalities deeply by introducing a soft-argmax layer and an IMU-bone layer in the network. DeepFuse does not require a given skeleton model *a priori*, and hence it can be well generalized to unknown subjects. Moreover, to make full use of the geometric-related frame data from multi-view images and preserve data primitiveness, we propose a new data representation: *multi-channel volume* for multi-view representation. Finally, we propose a new data augmentation technique, namely *Random Shut*, to enhance the generalization capability of our network for multi-channel volume data.

We test our fusion solution on TotalCapture [47] dataset, featuring synchronized camera data from 8 viewpoints, 13 body-worn IMUs and high-quality ground truth. Experimental results show that the proposed approach not only improves the estimation accuracy but also makes the two modalities of sensors mutually complementary. In addition, we test our vision-only network on a popular dataset Human3.6M [16] and achieve state-of-the-art results as well.

The main contributions of this work include:

- We propose a new vision-IMU data fusion technique namely **DeepFuse** for learning-based 3D human pose estimation, which deeply fuses data from the two kinds

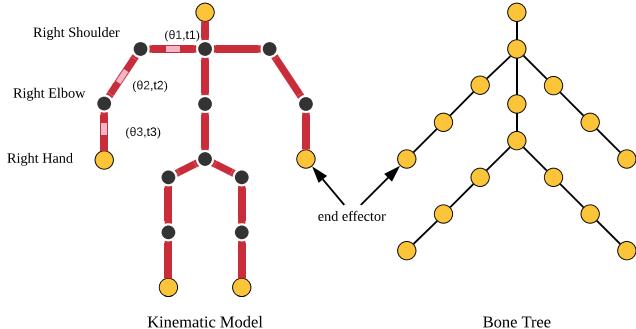


Figure 1: Illustration of kinematics. The position of end effector  $P_{rh}$ , right hand, can be derived through kinematic chain  $P_{rh} = R(\theta_3, t_3) \cdot R(\theta_2, t_2) \cdot R(\theta_1, t_1) \cdot P_{root}$ , where  $R(\theta, t)$  is the rotation matrix made from IMU orientation  $\theta$  and bone length  $t$ . See §2.2 for more details.

of sensors. Unlike previous works, the pre-defined skeleton model is not required in our method, making it well generalized to unknown users.

- We propose a new data format namely *multi-channel volume* with a corresponding data augmentation algorithm namely *Random Shut* to process multi-view images. This data format is able to preserve the geometric information of cameras and data primitiveness. Ablation study shows that Random Shut is effective in enhancing model generalization capability.
- To the best of our knowledge, this is the first work that applies soft-argmax layer in a fully 3D CNN network with volumetric input for human pose estimation. Our method outperforms state-of-the-art result by a large margin and we provide a detailed analysis to show effectiveness of 3D soft-argmax with volumetric representation by conducting rigorous experiments.

The remainder of this paper is organized as follows. In Section 2, we review the literature on human pose estimation. Section 3 details our method. Next, we present comparative study and ablation study with state-of-the-art works in Section 4. Limitations is discussed in Section 5. Finally, Section 6 concludes this paper.

## 2. Related work

There are mainly three types of human pose estimation methods: vision-based, IMU-based and hybrid approaches.

### 2.1. Vision-based human pose estimation

We can generally divide the vision-based tasks into 2D and 3D human pose estimation.

2D human pose estimation has been extensively explored by adopting either heatmap-based approaches [15, 32, 44, 45, 9, 6] or regression-based methods [18, 55, 3] that regress 2D image to joint coordinates directly. Specifically, Newell *et al.* [29] conduct deep conv-deconv hourglass models

which has been widely used as a backbone network by previous works [57, 2, 18, 8, 14].

Despite of the success towards 2D pose estimation, 3D pose estimation is yet under explored. Most methods for 3D pose estimation originate from 2D estimation work [20, 58, 54, 4]. Generally speaking, heatmap-based methods [23, 28, 40, 39, 30, 34, 46] show superior performance to that of direct regression work [21, 27, 42, 47]. Heatmap-based methods [23, 34, 39, 40] regress volumetric heatmaps from 2D image. Recently, many multi-view based methods [17, 43, 35, 10, 37, 19] try to get more effective and accurate information from different views. To reduce quantization error, some work [39, 23, 30, 24, 31] introduce soft-argmax layer to replace hard-argmax. The effectiveness of soft-argmax on networks with 2D images input has been extensively studied in [39]. However, the effectiveness of soft-argmax layer in a fully 3D CNN network with volumetric input for pose estimation has not been explored yet. We, therefore, conduct rigorous ablation study to demonstrate the effectiveness of soft-argmax layer in a fully 3D CNN network, shown in section 4.2.3.

### 2.2. IMU-based human pose estimation

IMUs measure bone orientation accurately when they are attached to the human body. According to state-of-the-art results, the mean measurement error of bone orientation produced by body-worn IMUs is about  $1.65^\circ$  [33], while that by the vision-based method is about  $12.1^\circ$  [51]. Consequently, they are widely used in applications wherein recovering bone orientation is sufficient [50, 33, 41, 38].

However, the IMU-based approach has several critical limitations when used to estimate human joint positions.

- A pre-defined skeleton model is required to solve a kinematic chain as shown in Fig 1 to recover joint positions. Therefore, manual calibration of the skeleton model is mandatory for each subject. Marcard *et al.* [53] combine IMUs with a skinned multi-person linear model [22] to recover the joint positions.
- Body skeleton is modeled as a tree-like structure in the kinematic model. Even if you have obtained correct limb lengths, the position of end effector node, such as hand, is determined by all its ancestor nodes, making estimation error accumulated dramatically.
- Last but not the least, IMU is unable to determine the position of subjects in world space. Although theoretically the subject's positions can be derived by a double integral of the acceleration data in the IMU [41, 38], such measurement error dramatically accumulates over time, making it almost impossible to calculate subject's positions.

### 2.3. Hybrid approach for human pose estimation

From the above, vision-based approaches are good at acquiring joint positions, but they are sensitive to body occlusion and illumination changes. IMU-based approaches, on the other hand, are capable of capturing accurate bone orientation stably, but fail to obtain joint positions. A hybrid solution that is able to fuse the two modalities effectively would have great potential.

Malleson *et al.* [26] propose a real-time optimization approach to fuse multi-view data and IMU data by combining position term, orientation term, pose prior term, and acceleration term. Particle-based optimization is used in [36] to constrain orientation cues from IMU and low-dimensional manifold images cues on an inverse kinematic model. Trumble *et al.* [47] propose a learning-based method to fuse volumetric data and IMU data in deep neural network. Bone orientations captured from IMU are converted to joint position by applying forward kinematics. And then, joint positions obtained from the two sources are fused at the very end of the network by fully connected layers. Consequently, the fusion layer makes limited contributions to vision tensors. Marcard *et al.* [51] achieve state-of-the-art result by solving a graph-based optimization that jointly optimizes vision data and IMU data on a SMPL model. They jointly optimize their model over all frames simultaneously, making it not applicable for a real-time system. Moreover, the overwhelming majority of current fusion work use pre-defined skeleton model, restricting the generalization capability of models to unknown subjects.

In terms of sensor fusion, tight coupling (fusion in data layer) shows overall better performance than loose coupling (fusion in result layer)[25]. We argue that for learning-based approach, the original data, instead of the estimation results, from the two data sources should be fused at the early stage of network so that the network could capture a deeper relationship of the two modalities to make them mutually complementary. Additionally, pre-defined skeleton model should not be introduced, as it would restrict the generalization capability of the model.

## 3. Proposed Solution

We define IMU-vision hybrid 3D human pose estimation as a learning-based two-stage regression problem. The network, namely *DeepFuse*, takes vision and IMU data as input, and regresses 3D human joint positions directly. As shown in Fig 2, the left part of the figure defines the estimation stage which infers 3D human pose from vision data only. The right part defines the refinement stage, introducing IMU-bone layer to refine the result from the previous stage. The whole network is trained in an end-to-end manner, making the two modalities mutually complementary

### 3.1. Pre-processing

The vision data from TotalCapture[48] is captured by 8 cameras. We use the provided binary matte images as input. To make full use of the geometric correlation of multi-view images and preserve data primitiveness, we propose a new data format called multi-channel volume and an accompanied data augmentation algorithm named Random Shut. The bone orientations read from IMU are transformed from local coordination to global coordination.

#### 3.1.1 Multi-channel volume

In terms of data representation in 3D human pose estimation, researchers either regard multi-view images as synchronized 2D images [39, 31] without considering multi-view geometry of cameras, or transform them into 3D volume [47, 46] by introducing probability visual hull (PVH). However, PVH does not well preserve data primitiveness since only one volume is generated by applying Bayes probability operator over all images. Accordingly, we propose a multi-channel volume format to overcome the limitations.

To build multi-channel volume for a single frame, we first define binary matte images  $I_1, I_2 \dots I_k$  from  $K$  cameras with intrinsic parameter  $M_k^{int}$  and extrinsic parameter  $M_k^{ext}$ . For each camera  $k$ , we initialize a volume  $V_k$  centred on the performer with resolution  $64 \times 64 \times 64$ . The voxel size is set to 35mm to make sure that all body parts are within the volume. The center position  $C$  of each voxel in volume  $V_k$  is transformed from world coordination to pixel coordination by camera parameter  $P_{x,y} = M_k^{int} \cdot M_k^{ext} \cdot C$ . The voxel value is set to 1 if its corresponding pixel in image  $I_k$  is occupied and to 0 otherwise. Instead of fusing the  $K$  volumes into one volume by applying summation or multiplication [47, 46], we simply regard each volume as a single channel of input to preserve original information as much as possible. Finally, the shape of the multi-channel volume is  $8 \times 64 \times 64 \times 64$ , serving as the vision input of our network. Fig 3 shows a sample multi-channel volume and its respective matte images.

#### 3.1.2 Data augmentation

In order to prevent overfitting, the volumes are augmented by performing a random rotation around the vertical axis within the range  $[-\pi, \pi]$ . Additionally, as shown in section 4.2.2, the estimation accuracy decreases when the subject is not captured by all the 8 cameras at the same time. In order to make our model well adapted to this situation, chances are that a random volume channel is set to zero to augment training data. We set the chance of randomly 'shutting down' a camera to 20% and name it as Random Shut. This data augmentation algorithm is proved to be effective for multi-view data in ablation study 4.2.2.

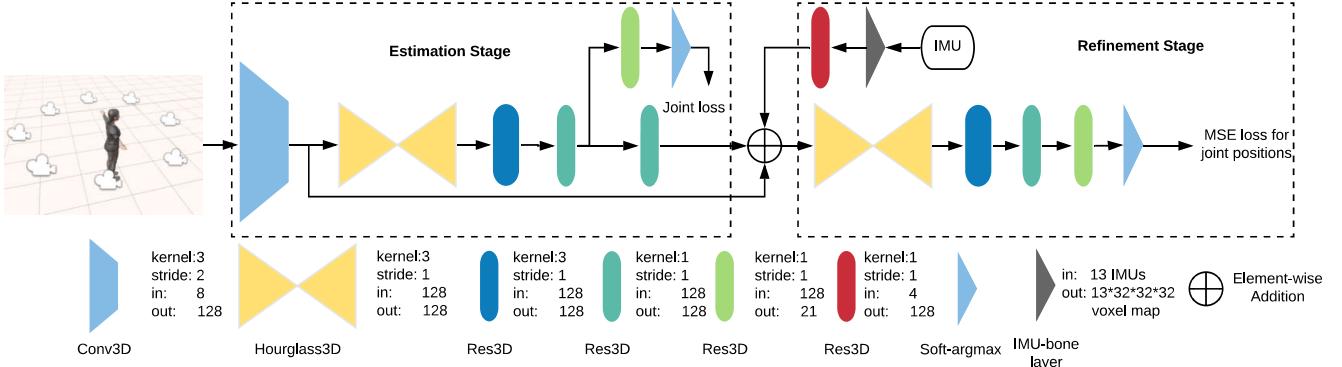


Figure 2: To simplify the illustration, all the 3D modules are visualized with 2D shapes. In the estimation stage, vision data is first down-sampled and passes through the hourglass network, Residual network 3D (Res3D) and soft-argmax layer. The first mean square error (MSE) loss between estimation result and ground truth is computed at the end of this stage. In the refinement stage, bone orientations from IMUs are transformed to  $13 \times 32 \times 32 \times 32$  volume by IMU-bone layer, which are then concatenated with vision volume and heatmap volume. See for more details.

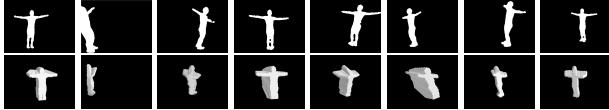


Figure 3: Sample multi-channel volume of a single frame. The first row is the given matte data and the second row is its respective channel of volume (See §3.1.1).

Method	Error (mm)
Video Inertial Poser (VIP) [51]	26.0†
Frame-by-Frame Optimization [26]	62.0
FC IMU+3D PVH [47]	70.0
<b>DeepFuse</b>	<b>28.9</b>

Table 1: Comparison results regarding mean joint error on TotalCapture dataset. † indicates sequence-based work. See §4.1.1 for detail. (Best in **bold**; same for other tables).

### 3.1.3 IMU orientation

According to [47], IMUs are assumed rigid attached to human bones. The orientation data for sensor  $k$  is measured in local frames and employed in quaternion representation (as in: [12, 7, 1]), noted as  ${}^{local}q_k$ . By multiplying wearing offset  ${}^{wear}q_k$  and local-global transform quaternion  ${}^{global}q_k$ , the bone orientation in global frame  ${}^{global}\hat{q}_k$  is calculated as:

$${}^{global}q_k = {}^{local}q_k \otimes {}^{global}q_k \quad (1)$$

$${}^{global}\hat{q}_k = {}^{wear}q_k^* \otimes {}^{global}q_k \quad (2)$$

where “\*” denote the quaternion conjugate.

The bone orientations in global frame  ${}^{global}\hat{q}_k$  are then transformed into IMU-bone layer in the network as discussed in section 3.2.2.

## 3.2. Network structure

As shown in Fig 2, DeepFuse consists of an estimation stage and a refinement stage. The backbone network used is Hourglass Network [29]. The network takes multi-channel volume as vision input, so we modify original Hourglass Network to the 3D version with 3D CNN. The left part of Fig 2 represents the estimation stage which takes vision data only as input and outputs 3D voxel heatmaps of each joint. Considering the large GPU memory consumption of 3D CNN, the volume input resolution for one channel is  $64 \times 64 \times 64$  and the voxel heatmap is  $32 \times 32 \times 32$ . By adding a soft-argmax layer to the end of the hourglass network, the 3D positions of each joint can be directly regressed from voxel heatmap. Soft-argmax is differentiable so that the estimated positions are able to further propagate to generate IMU-bone layer. Thus, the entire network can be trained in an end-to-end manner.

The right part is the refinement stage, taking IMU data and output from the previous stage as input. Specifically, the IMU data and estimated joint positions from the last stage consist the IMU-bone layers which turn quaternion data into a multi-channel volume. Thus, IMU-bones layers, voxel heatmap and original vision data, can be concatenated in the same feature space. In the following stack of hourglass network, IMU data and vision data are fused implicitly to produce a set of refined voxel heatmaps. A soft-argmax layer is appended to the last layer as well, from which the final refined joint positions are estimated.

### 3.2.1 Soft-argmax layer

Heatmap-based approaches are proved to be effective in both human pose and hand pose estimation [29, 14]. The voxel heatmap is discrete, while the joint positions are continuous in world coordination. So if we assume that volume

Protocol 1	Direct	Discuss	Eat	Greet	Phone	Photo	Pose	Purcha.	Sit	SitD	Smoke	Wait	WalkD	Walk	WalkT	Avg.
Fang2018 w/PA[11]	38.2	41.7	43.7	44.9	48.5	55.3	40.2	38.2	54.5	64.4	47.2	44.3	47.3	36.7	41.7	45.7
Sun2018 w/PA[39]	40.9	41.4	45.0	45.2	42.1	37.6	41.1	52.0	71.4	42.5	47.4	41.6	32.0	42.6	36.9	44.1
Chen2019 w/PA[5]	36.9	39.3	40.5	41.2	42.0	34.9	38.0	51.2	67.5	42.1	42.5	37.5	30.6	40.2	34.2	41.6
Multi-View Martinez2017 w/PA[27]×	39.5	43.2	46.4	47.0	51.0	56.0	41.4	40.6	56.5	69.4	49.2	45.0	49.5	38.0	43.1	47.7
Tome2018 w/PA[43]×	38.2	40.2	38.8	41.7	44.5	54.9	34.8	35.0	52.9	75.7	43.3	46.3	44.7	35.7	37.5	44.6
<i>Ours</i> w/o PA † ×	<b>18.7</b>	<b>20.7</b>	<b>22.5</b>	<b>24.5</b>	<b>28.3</b>	40.1	<b>22.7</b>	<b>23.1</b>	<b>26.0</b>	<b>39.9</b>	<b>33.8</b>	<b>22.9</b>	<b>35.0</b>	<b>20.9</b>	<b>21.3</b>	<b>26.9</b>
<i>Ours</i> w/PA † ×	<b>9.0</b>	<b>9.4</b>	<b>11.2</b>	<b>13.5</b>	<b>14.0</b>	<b>22.1</b>	<b>11.6</b>	<b>12.0</b>	<b>11.8</b>	<b>20.8</b>	<b>14.7</b>	<b>10.6</b>	<b>20.4</b>	<b>10.4</b>	<b>10.7</b>	<b>13.4</b>
Protocol 2	Direct	Discuss	Eat	Greet	Phone	Photo	Pose	Purcha.	Sit	SitD	Smoke	Wait	WalkD	Walk	WalkT	Avg.
Trumble2017 [47] †	92.7	85.9	72.3	93.2	86.2	101.2	75.1	78.0	83.5	94.8	85.8	82.0	114.6	94.9	79.7	87.3
Trumble2018 [46] † ×	41.7	43.2	52.9	70.0	64.9	83.0	57.2	63.5	61.0	95.0	70.0	62.3	66.2	53.7	52.4	62.5
Multi-View Martinez2017 [27][43]×	46.5	48.6	54.0	51.5	67.5	70.7	48.5	49.1	69.8	79.4	57.8	53.1	56.7	42.2	45.4	57.0
Pavlakos2017[35]×	41.2	49.2	42.8	43.4	55.6	46.9	40.3	63.7	97.6	119.0	52.1	42.7	51.9	41.8	39.4	56.9
Tome2018 [43]×	43.3	49.6	42.0	48.8	51.1	64.3	40.3	43.3	66.0	95.2	50.2	52.2	51.1	43.9	45.3	52.8
Kocabas2019-FS[19]×	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	51.8
Kadkhodamohammadi2018[17]×	39.4	46.9	41.0	42.7	53.6	54.8	41.4	50.0	59.9	78.8	49.8	46.2	51.1	40.5	41.0	49.1
<i>Ours</i> † ×	<b>26.8</b>	<b>32.0</b>	<b>25.6</b>	<b>52.1</b>	<b>33.3</b>	<b>42.3</b>	<b>25.8</b>	<b>25.9</b>	<b>40.5</b>	<b>76.6</b>	<b>39.1</b>	<b>54.5</b>	<b>35.9</b>	<b>25.1</b>	<b>24.2</b>	<b>37.5</b>

Table 2: Comparison results regarding mean joint error following protocol 1 and 2 of Human3.6M with 17 keypoints. † use provided matte data. × use multi-view images. (FS: fully supervised baseline). See §4.1.2 for details.

length is 2000mm and heatmap resolution is  $32 \times 32 \times 32$ , the average estimation error within a single voxel is about  $2000/32/2 = 31.25\text{mm}$  at least. So even if we obtained a perfect voxel heatmap, the accuracy of joint positions in world coordination would be far from satisfactory by simply picking the largest voxel in a voxel heatmap.

To overcome the limitation of low resolution of voxel heatmap, we introduce soft-argmax layer into our network. Instead of simply picking the largest voxel as joint position, the soft-argmax layer is able to learn a weighted average of multiple voxels to predict joint positions.

Specifically, soft-argmax shares similar idea with *softmax* algorithm. Softmax value of voxel  $x_i$  is defined as:

$$f_{\text{softmax}}(x_i) = \frac{e^{\theta x_i}}{\sum_j^N e^{\theta x_j}} \quad (3)$$

The sum of softmax values equals to 1. Thus, the coordinate of max value among all voxels is the sum of softmax values multiplied by indices along each axis:

$$f(x) = \sum_i \frac{e^{\theta x_i}}{\sum_j^N e^{\theta x_j}} i \quad (4)$$

where  $i$  is the x-,y-,z-coordinate of voxel  $x_i$  for each axis and  $N$  is the total voxel number in the volume. The larger  $\theta$  will enlarge the voxels with big value and lower the smaller ones, which means if  $\theta$  is large enough, soft-argmax will return the coordinate of voxel with maximum value. However, we expect that the joint position should be the coordinate of the weighted average of several large voxels. We set  $\theta = 3$  empirically (see ablation study in section 4.2.3).

### 3.2.2 IMU-bone layer and sensor fusion

As explained in section 2.2, IMU sensor measures bone orientation only. Thus, the pre-defined skeleton model was introduced to estimate joint positions as [51, 47, 52] did in their work. We claim that the pre-defined skeleton model should not be used as prior knowledge in order to make the estimation model well generalized to unknown subjects.

Therefore, the challenge of sensor fusing lies in that how to use measured bone orientation from IMU to refine joint position estimation result from vision.

Existing learning-based fusion work [47] tries to fuse the estimation results from two modalities instead of origin inputs through a dense layer at the end of the network. As a result, the original data from two sensors does not propagate to its counterpart and fusion is relatively shallow. So we are going to fuse the original input of two modalities at the early stage of the network for a deeper fusion.

According to the skeleton model, one limb  $L$  consists of two joints  $[J_1, J_2]$  and one piece of bone. If the position of one joint, say  $J_1$ , and bone orientation  $q_b$  are given, a ray  $f_{\text{ray}}(J_1, q_b)$  can be cast from  $J_1$  along  $q_b$  direction. Thus,  $J_2$  must be located somewhere along the ray. So if we can obtain the estimated joint positions  $J_{\text{est}}$  from estimation stage and the bone orientation  $q_b$  from IMU, a cluster of rays can be obtained respectively to describe the possible locations of joints. Since  $J_{\text{est}}$  is the estimated result, the generated rays are not 100% accurate. The rays are transformed to directed cylinders by introducing a radius  $r$  around the rays. Finally, we define a channel of volume  $V_i = f_{\text{cylinder}}(J_{\text{est}}[i], q_b[i], r)$  for each IMU  $i$ . The voxels are set to 1 if occupied by the 'bone cylinder' and to 0 otherwise. Finally, we stack these volumes together to form IMU-bone layers.

Since we have volumetric representation for original vision data, voxel heatmap for each joint, and volumes for IMU-bone layer at this stage, the three batches of volumes can be concatenated together and then passed to next stage to refine the result from estimation stage. In this way, the original data from two modalities are deeply fused by a new stack of hourglass network in the refinement stage. Similarly, a soft-argmax layer is appended to the last layer to make final estimation of joint positions.

### 3.2.3 Training target

By introducing soft-argmax layer, joint positions can be directly recovered. So there is no need to involve in heatmap

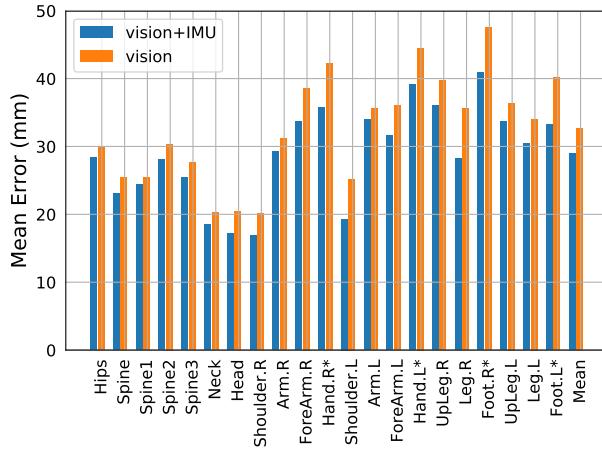


Figure 4: Per-joint mean error on test set. \* indicates right/left foot and hand joints. See §4.2.1 for details.

Methods	Error (mm)
Vision-IMU w/ RS (proposed)	<b>28.9</b>
Vision-only w/ RS	32.7
Vision-IMU w/o RS	<b>32.4</b>
Vision-only w/o RS	35.9

Table 3: Mean joint error results for 4 different experiment settings. RS is short for Random Shut. Vision-IMU w/ RS is identical to the proposed *DeepFuse* (See §4.2.1).

loss as the voxel heatmap can be learned implicitly by the mean squared error between the estimated joint positions and ground truth positions in world coordination. Specifically, the loss function  $L$  is the sum of estimation stage loss  $L_{est}$  and refinement stage loss  $L_{ref}$ :

$$\begin{aligned} L &= L_{est} + L_{ref} \\ &= \sum_i^K |J_{est}^i - J_{gt}^i|^2 + \sum_i^K |J_{ref}^i - J_{gt}^i|^2 \end{aligned} \quad (5)$$

where  $J_{est}^i$  and  $J_{ref}^i$  are the estimated joint positions for  $i^{th}$  joint in estimation stage and refinement stage, respectively.  $J_{gt}^i$  are the ground truth.  $K$  is the #joints to be estimated.

## 4. Experimental Results

In our experiments, RMSProp optimizer is used for training. The learning rate is initially set to be 1e-5 and decays 0.2 every 5 epochs. Our system is able to run at 25 Hz with a single NVIDIA 1080Ti GPU.

### 4.1. Comparative study

In this section, we present the comparative result of *DeepFuse* with recent works on two public 3D human pose estimation datasets: TotalCapure [47] and Human3.6M [16]. The TotalCapture dataset features 1.9M video frames

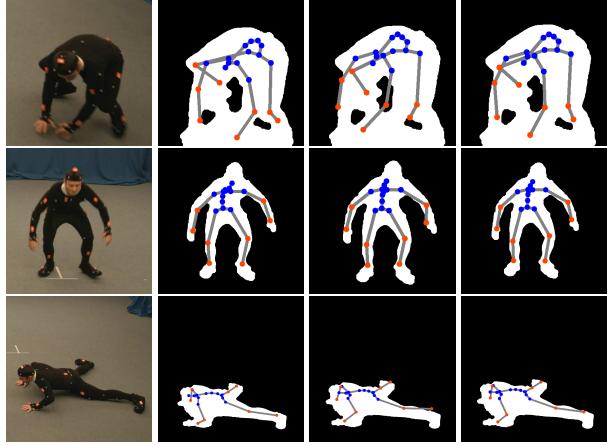


Figure 5: Left to right: camera, ground truth, vision-only estimation, fusion estimation. See §4.2.1 for details.

from 8 cameras performed by 5 subjects and synchronized orientation data from 13 body-worn IMUs. Human3.6M is a more popular 3D human pose estimation dataset with only vision data as input.

#### 4.1.1 TotalCapture evaluation

TotalCapture [47] is the only dataset including synchronized body-worn IMU data and multi-view video frames with high-quality ground truth. There are three pieces of work [47, 26, 51] evaluated on this dataset by using both vision and IMU data. Specifically, Malleson *et al.* [26] optimizes the two modalities in a frame-by-frame manner while Marcard *et al.* [51] optimizes the whole video sequence simultaneously. Learning-based method [47] uses fully connected layer to fuse the IMU data and 3D vision data.

As shown in Table 1, *DeepFuse* outperforms [47, 26] by a large margin and shows close performance with [51]. However, pre-defined skeleton model is not employed in our method, but it was used by all the other three methods. Consequently, *DeepFuse* is more friendly to unknown subjects. Moreover, Video Inertial Poser (VIP) [51] achieved the lowest estimation error by optimizing over all frames of a given video sequence and hence it is not applicable for real-time estimation. *DeepFuse* and the other two methods are frame-based estimations without such limitation.

To sum up, our method achieves the state-of-the-art result in terms of real-time frame-based 3D human pose estimation by fusing IMU and vision data on TotalCapture, showing that even if there is no pre-defined skeleton model as input, vision data and IMU data can still be fused in a complementary way to produce better fusion result.

#### 4.1.2 Human3.6M evaluation

To validate the performance of our proposed multi-channel volume data representation and soft-argmax on volumetric

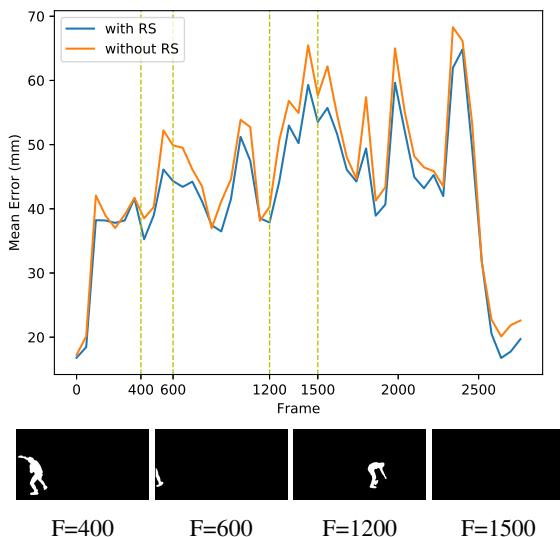


Figure 6: Per-frame mean joint error of test sequence S4 F3. The four matte images at bottom are captured by camera 4 at frame 400, 600, 1200 and 1500. Partially-captured frames, *e.g.* frame 600 and 1500, show inferior performances. *RS* for *Random Shut*. *F* denotes frame number. (See §4.2.2).

data, we remove the IMU-bone layer of our network and test it on Human3.6M dataset [16]. It consists of 3.6M frames captured from 11 subjects with 4 synchronized cameras. Followed by the setting of [39], protocol 1 uses subjects (S1, S5, S6, S7, S8, S9) for training and S11 for test. The measured result, mean per joint position error ( MPJPE ), is aligned by Procrustes Analysis ( PA MPJPE ). Protocol 2 uses subjects (S1, S5, S6, S7, S8) for training and subjects (S9, S11) for test without PA. To remove data redundancy, only every 5th frames in training sequences and every 64th frames in test sequences are used. Meanwhile, we use the provided foreground matte images from all the 4 cameras as input. Because the number of cameras is only 4 and subject always move within area captured by all cameras, Random Shut introduces more noise on this dataset in our experiment, which is not used in training as a result.

From table 2, our method significantly outperforms other methods by large margins including singles-view methods and multi-view methods. Specifically, it improves the SOTA [43] by **31.2mm** (relative 70% lower than [43]) under protocol 1, and by **11.6mm** (relative 24.0% lower than the state-of-the-art method [19]) under protocol 2, and it also is clearly ahead by over 40% as the same matte data used in [46]. By observing the failure cases, we find that the matte data of 3 actions, Greeting, Sitting Down and Waiting, are incomplete due to truncations, leading to unexpected high error. After remove these there actions, our result reaches 32.5mm. Furthermore, we will show the qualitative examples in the supplementary materials.

Training	Full Frames	Partial Frames	Overall
with RS	18.9	34.5	28.9
without RS	20.0	39.2	32.4

Table 4: Mean joint error of our proposed method trained *w.* or *w/o.* Random Shut (RS). Full frames are frames that are captured by all the 8 cameras while partial frames are not. All errors are measured in millimeter (mm) (See §4.2.2).

Finally, recent methods with volumetric input show overall better performance than that with 2D image input, showing the advantage of volumetric data representation. Also, our method outperforms other volume-based method by a large margin, showing the effectiveness of soft-argmax layer on volumetric data for 3D human pose estimation.

## 4.2. Ablation study

In this section, we try to answer the following three questions: (1) To what extent can IMU-bone layer contribute to the final estimation? (2) Can the proposed data augmentation algorithm, Random Shut, improve the generalization capability of our model? And (3) To what extent and why 3D soft-argmax over volumetric data representation improve the estimation accuracy? All the experiments of ablation study are evaluated on the **TotalCapture** dataset[48].

### 4.2.1 Sensor fusion

The first study explores the effectiveness of sensor fusion. Results of four training strategies are listed in Table 3. To alleviate the possible influence of Random Shut (RS) on this ablation study, we perform comparison experiment between vision-IMU network and vision-only network *w.* or *w/o.* RS, respectively. Both of the vision-IMU networks outperform their vision-only counterparts, supporting the effectiveness of our data fusion solution quantitatively.

In addition, we want to find out how IMU data influence the estimation after refinement. We plot the per-joint estimation error on test set. As shown in Fig 4, the joints around limbs including foot and hand show more improvement by fusing IMU sensors compared to other joints. The main reason is that the IMU-bone layer constructs cylinder volume for each bone and it simulates volume around limbs much better than that around torso due to visual similarity. This argument is also supported by the qualitative result in Fig.5. As can be observed, fusion result is superior to vision-only result. The joints around limbs show better refinement than that around torso, especially under heavy self-occlusion.

As discussed before, the measurement of IMU sensors is more stable than that of vision sensors. Therefore, in supplementary material, we show that fusion approach shows better sequential stability than its vision-only counterpart.

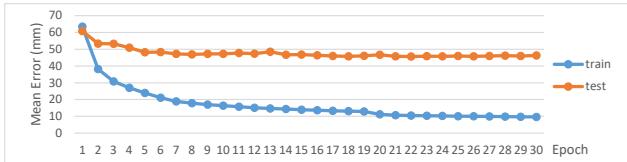


Figure 7: Mean error over epoch

Baselines	Hard-argmax	Soft-argmax	Direct regression (FC)
Mean Error (mm)	39.7	<b>32.7</b>	45.2

Table 5: Mean joint error results for three kinds of output layers. FC stands for fully connected layer (See §4.2.3).

#### 4.2.2 Random Shut

The second study shows the motivation and effectiveness of the proposed data augmentation approach, Random Shut (RS), for multi-view data. Fig 6 demonstrates the per-frame estimation accuracy on a test sequence. We find that the estimation error increases when subject is partially-captured (Frame 600 and 1500), which is also proved by statistic data shown in Table 4. Motivated by this finding, Random Shut simulates the situation when subject randomly missed by a certain camera during training. As demonstrated in Table 4 and in Fig. 6, we conclude that RS shows improvement on generalization capability of the model for multi-view data, especially for the partially-captured frames.

#### 4.2.3 Soft-argmax and volumetric representation

The last study aims to validate the effectiveness of 3D soft-argmax layer and volumetric representation. Although soft-argmax layer has already been used in human pose estimation by several works [39, 23, 30], the effectiveness of this layer has not yet been evaluated in a fully 3D CNN-based network with volumetric input.

To show the effectiveness of **3D soft-argmax**, we compare it with two widely used post-processing techniques: **hard-argmax** which directly picks highest voxel position, and **direct regression** which does regression via a dense layer. For fair comparison, we evaluate the three kinds of methods on the proposed fully 3D CNN-based network with volumetric data as input only. The comparative results in Table 5 show that soft-argmax significantly outperforms the other approaches. The main reason is that the low resolution of voxel heatmap limits the performance of hard-argmax as stated in section 3.2.1 and the dense layer is not well capable of learning argmax operator which is highly non-linear.

To show the effectiveness of **volumetric representation**, we first have following finding in Table 2 that our volumetric representation achieves mean error of **37.5mm**, an improvement of **23.6%** and **28.9%** compared to its 2D soft-argmax multiview counterparts 49.1mm[17] and 52.8mm[43], showing the advantage of volumetric representation in 3D human pose estimation. Second, our model converges very fast with 3D soft-argmax. As shown in

$\theta$	1	2	3	5	10	100
Error (mm)	33.8	33.4	<b>32.7</b>	34.2	38.9	41.0

Table 6: Different  $\theta$  of soft-argmax and their respective estimation errors in the vision-only network (See §4.2.3).

Fig. 7, we achieve state-of-the-art result only after the first epoch, indicating that fully 3D CNN-based network would not waste its capability on learning unnecessary space mapping, which explains the effectiveness of volumetric representation.

Meanwhile, the model converges too fast, indicating potential severe overfitting. We therefore explore how to tune the parameter  $\theta$  in the soft-argmax Equation 4, which may influence the convergence speed. As discussed in Sec. 3.2.1, soft-argmax pays more attention to large voxel when  $\theta$  enlarges. If  $\theta$  is too large, soft-argmax will degenerate to hard-argmax. If  $\theta$  is too small, too many voxels contribute to the final result, which may lead to overfitting. So we list some possible values of  $\theta$  and their respective estimation errors in Table 6. In our experiment,  $\theta = 3$  achieves the best performance in our experiment setting. We believe the above findings of using soft-argmax in 3D space are novel and would be beneficial for future research.

## 5. Limitation and future work

One limitation is that, like many multi-view methods, the proposed multi-channel volume is biased to camera configuration. To make it fair in more general configurations, the triangulation should be trainable instead of fixed, making model adaptive. Another one is that performance of our method depends on the quality of foreground silhouette. The popular segmentation networks, like maskrcnn[13], show low quality of segmentation near human edge so we use the given silhouette ground truth. In the future work, we will try to make this step into an end-to-end pipeline for simplifying the input representation and improve the silhouette quality simultaneously.

## 6. Conclusion

In this paper, we propose an IMU-aware network to fuse IMU data and multi-view images for 3D human pose estimation. To fully utilize the multi-view geometric information, we re-project it into a multi-channel volume format and apply Random Shut for data augmentation. To deeply fuse IMU orientation and multiple views, we then propose an IMU-bone layer to transform the original data from two modalities into a same feature space at early stage of network. Rigorous ablation shows the effectiveness of the multi-channel volume, Random Shut and IMU-bone layer. Finally, our method achieves the state-of-the-art performance on both TotalCapture and Human3.6M dataset with real-time capability.

## References

[1] E. R. Bachmann, I. Duman, U. Usta, R. B. McGhee, X. Yun, and M. Zyda. Orientation tracking for humans and robots using inertial sensors. In *Proceedings 1999 IEEE International Symposium on Computational Intelligence in Robotics and Automation. CIRA'99 (Cat. No. 99EX375)*, pages 187–194. IEEE, 1999.

[2] A. Bulat and G. Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *European Conference on Computer Vision*, pages 717–732. Springer, 2016.

[3] J. Carreira, P. Agrawal, K. Fragiadaki, and J. Malik. Human pose estimation with iterative error feedback. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4733–4742, 2016.

[4] C.-H. Chen and D. Ramanan. 3d human pose estimation= 2d pose estimation+ matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7035–7043, 2017.

[5] X. Chen, K.-Y. Lin, W. Liu, C. Qian, and L. Lin. Weakly-supervised discovery of geometry-aware representation for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10895–10904, 2019.

[6] Y. A. Chen, X. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *the Neural Information Processing Systems*, page 30413048, 2014.

[7] D. Choukroun. *Novel methods for attitude determination using vector observations*. Technion-Israel Institute of Technology, Faculty of Aerospace Engineering, 2003.

[8] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang. Multi-context attention for human pose estimation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1831–1840, 2017.

[9] M. Dantone, J. Gall, C. Leistner, and L. Van Gool. Human pose estimation using body parts dependent joint regressors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3041–3048, 2013.

[10] J. Dong, W. Jiang, Q. Huang, H. Bao, and X. Zhou. Fast and robust multi-person 3d pose estimation from multiple views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7792–7801, 2019.

[11] H.-S. Fang, Y. Xu, W. Wang, X. Liu, and S.-C. Zhu. Learning pose grammar to encode human body configuration for 3d pose estimation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[12] D. Gebre-Egziabher, R. C. Hayward, and J. D. Powell. Design of multi-sensor attitude determination systems. *IEEE Transactions on aerospace and electronic systems*, 40(2):627–649, 2004.

[13] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[14] F. Huang, A. Zeng, M. Liu, J. Qin, and Q. Xu. Structure-aware 3d hourglass network for hand pose estimation from single depth image. In *British Machine Vision Conference*, page 289, 2018.

[15] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepcut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*, pages 34–50. Springer, 2016.

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

[17] A. Kadkhodamohammadi and N. Paday. A generalizable approach for multi-view 3d human pose regression. *CoRR*, abs/1804.10462, 2018.

[18] L. Ke, M.-C. Chang, H. Qi, and S. Lyu. Multi-scale structure-aware network for human pose estimation. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[19] M. Kocabas, S. Karagoz, and E. Akbas. Self-supervised learning of 3d human pose using multi-view geometry. *arXiv preprint arXiv:1903.02330*, 2019.

[20] S. Li and A. B. Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *Asian Conference on Computer Vision*, pages 332–347. Springer, 2014.

[21] M. Lin, L. Lin, X. Liang, K. Wang, and H. Cheng. Recurrent 3d pose sequence machines. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 5543–5552. IEEE, 2017.

[22] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):248, 2015.

[23] D. C. Luvizon, D. Picard, and H. Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5137–5146, 2018.

[24] D. C. Luvizon, H. Tabia, and D. Picard. Human pose regression by combining indirect part detection and contextual information. *arXiv preprint arXiv:1710.02322*, 2017.

[25] M. Ma, C. Sun, and X. Chen. Deep coupling autoencoder for fault diagnosis with multimodal sensory data. *IEEE Transactions on Industrial Informatics*, 14(3):1137–1145, March 2018.

[26] C. Malleson, A. Gilbert, M. Trumble, J. Collomosse, A. Hilton, and M. Volino. Real-time full-body motion capture from video and imus. In *3D Vision (3DV), 2017 International Conference on*, pages 449–457. IEEE, 2017.

[27] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *International Conference on Computer Vision*, volume 1, page 5, 2017.

[28] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4):44, 2017.

[29] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016.

[30] A. Nibali, Z. He, S. Morgan, and L. Prendergast. 3d human pose estimation with 2d marginal heatmaps. *arXiv preprint arXiv:1806.01484*, 2018.

[31] A. Nibali, Z. He, S. Morgan, and L. Prendergast. Numerical coordinate regression with convolutional neural networks. *arXiv preprint arXiv:1801.07372*, 2018.

[32] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy. Towards accurate multi-person pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4903–4911, 2017.

[33] M. Paulich, M. Schepers, N. Rudigkeit, and G. Bellusci. Xsens mtw awinda: Miniature wireless inertial-magnetic motion tracker for highly accurate 3d kinematic applications.

[34] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. *IEEE Conference on Computer Vision and Pattern Recognition*, abs/1611.07828, 2017.

[35] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Harvesting multiple views for marker-less 3d human pose annotations. *Conference on Computer Vision and Pattern Recognition*, page 12531262, 2017.

[36] Pons-Moll, G. Baak, A., Gall, J., Leal-Taixe, L., Muller, M., Seidel, H.P., Rosenhahn, and B. Outdoor human motion capture using inverse kinematics and von mises-fisher sampling, 2011.

[37] H. Rhodin, J. Spörri, I. Katircioglu, V. Constantin, F. Meyer, E. Müller, M. Salzmann, and P. Fua. Learning monocular 3d human pose estimation from multi-view images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8437–8446, 2018.

[38] R. Slyper, Hodgins, and J.K. Action capture with accelerometers. In *Proceedings of the 2008 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 193–199, 2008.

[39] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 529–545, 2018.

[40] I. Srndi, T. Linder, K. O. Arras, and B. Leibe. Synthetic occlusion augmentation with volumetric heatmaps for the 2018 eccv posetrack challenge on 3d human pose estimation, 2018.

[41] Tautges, J., Zinke, A. Krger, B., Baumann, J., Weber, A., Helten, T., Mller, M., Seidel, H.P., and B. Eberhardt. Motion reconstruction using sparse accelerometer data. *ACM transactions on graphics (TOG)*, 30(3):18, 2011.

[42] B. Tekin, P. Marquez Neila, M. Salzmann, and P. Fua. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In *International Conference on Computer Vision (ICCV)*, number EPFL-CONF-230311, 2017.

[43] D. Tome, M. Toso, L. Agapito, and C. Russell. Rethinking pose in 3d: Multi-stage refinement and recovery for markerless motion capture. In *2018 International Conference on 3D Vision (3DV)*, pages 474–483. IEEE, 2018.

[44] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. Efficient object localization using convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 648–656, 2015.

[45] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in neural information processing systems*, pages 1799–1807, 2014.

[46] M. Trumble, A. Gilbert, A. Hilton, and J. Collomosse. Deep autoencoder for combined human pose estimation and body model upscaling. In *European conference on computer vision (ECCV18)*, 2018.

[47] M. Trumble, A. Gilbert, C. Malleson, A. Hilton, and J. Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *Proceedings of 28th British Machine Vision Conference*, pages 1–13, 2017.

[48] M. Trumble, A. Gilbert, C. Malleson, A. Hilton, and J. Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *2017 British Machine Vision Conference (BMVC)*, 2017.

[49] Vicon. Vicon motion systems ltd. <http://www.vicon.com/>.

[50] D. Vlasic, R. Adelsberger, G. Vannucci, J. Barnwell, M. Gross, W. Matusik, and J. Popovi. Practical motion capture in everyday surroundings. *ACM transactions on graphics (TOG)*, 26(3):35, 2007.

[51] T. von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, 2018.

[52] T. von Marcard, G. Pons-Moll, and B. Rosenhahn. Human pose estimation from video and imus. *IEEE transactions on pattern analysis and machine intelligence*, 38(8):1533–1547, 2016.

[53] T. von Marcard, B. Rosenhahn, M. J. Black, and G. Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer Graphics Forum*, volume 36, pages 349–360. Wiley Online Library, 2017.

[54] C. Wang, Y. Wang, Z. Lin, A. L. Yuille, and W. Gao. Robust estimation of 3d human poses from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2361–2368, 2014.

[55] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

[56] Xsens. Xsens motion technologies. <http://www.xsens.com/>.

[57] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang. Learning feature pyramids for human pose estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 2, 2017.

[58] H. Yasin, U. Iqbal, B. Kruger, A. Weber, and J. Gall. A dual-source approach for 3d pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4948–4956, 2016.