C3P: Cross-domain Pose Prior Propagation for Weakly Supervised 3D Human Pose Estimation

Cunlin Wu[1], Yang Xiao[1†], Boshen Zhang[2], Mingyang Zhang[1], Zhiguo Cao[1], and Joey Tianyi Zhou[3]

[1] Key Laboratory of Image Processing and Intelligent Control, Ministry of Education, School of AIA, Huazhong University of Science and Technology, China
{cunlin_wu,Yang_Xiao,izmy,zgcao}@hust.edu.cn
[2] YouTu Lab, Tencent
boshenzhang@tencent.com
[3] A*STAR Centre for Frontier AI Reasearch (CFAR), Singapore.
zhouty@ihpc.a-star.edu.sg

Abstract. This paper first proposes and solves weakly supervised 3D human pose estimation (HPE) problem in point cloud, via propagating the pose prior within unlabelled RGB-point cloud sequence to 3D domain. Our approach termed C3P does not require any labor-consuming 3D keypoint annotation for training. To this end, we propose to transfer 2D HPE annotation information within the existing large-scale RGB datasets (e.g., MS COCO) to 3D task, using unlabelled RGB-point cloud sequence easy to acquire for linking 2D and 3D domains. The self-supervised 3D HPE clues within point cloud sequence are also exploited, concerning spatial-temporal constraints on human body symmetry, skeleton length and joints' motion. And, a refined point set network structure for weakly supervised 3D HPE is proposed in encoder-decoder manner. The experiments on CMU Panoptic and ITOP datasets demonstrate that, our method can achieve the comparable results to the 3D fully supervised state-of-the-art counterparts. When large-scale unlabelled data (e.g., NTU RGB+D 60) is used, our approach can even outperform them under the more challenging cross-setup test setting. The source code is released at https://github.com/wucunlin/C3P for research use only.

Keywords: 3D human pose estimation, weak supervision, RGB-point cloud sequence, spatial-temporal constraints

1 Introduction

3D human pose estimation (HPE) in depth data (e.g., depth map or point cloud) is of wide-range applications towards human action recognition [22, 1, 23], human-robot interaction [38], virtual [29], etc. With the introduction of deep learning technologies (e.g., CNN [36, 12] or PointNet [30, 31]), 3D HPE's

†Yang Xiao is corresponding author (Yang_Xiao@hust.edu.cn).

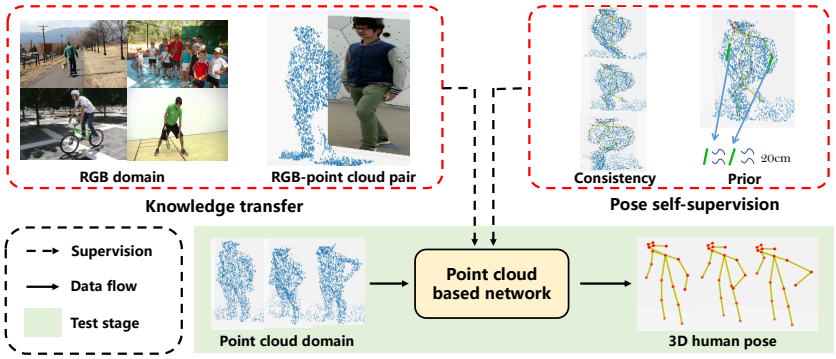


Fig. 1. The main research idea of our weakly supervised 3D human pose estimation approach in point cloud. RGB-point cloud pairs bridge 2D and 3D domain. 3D priors of human pose in point cloud sequences are also utilized as self-supervision signals. For test, we only need point cloud as input, and output 3D human pose in world coordinate.

performance has been enhanced remarkably in fully supervised learning manner. However, deep network's data-hungry property leads to the high demand on 3D pose annotation both on quality and quantity, which is essentially labor and time consuming. Nevertheless, the existing annotated 3D HPE datasets are generally of relative small size. For example, ITOP [11] only involves 50K samples from 20 subjects in laboratory setting. While, as the 2D RGB counterpart MS COCO [21] contains over 200K samples from 250K subjects under in the wild conditions. Accordingly, the existing RGB-based 2D HPE approaches [4, 43, 42, 28, 5] are generally of stronger generality. Thus, we raise the question that *whether the rich 2D annotation information within RGB domain can be transferred to depth domain for facilitating 3D HPE*, which has not been well concerned before.

Due to the emergency of low-cost RGB-D cameras (e.g., MS Kinect [25, 24]), 3D human pose's unlabelled RGB-D pair sequence can be easily acquired to link 2D and 3D domains. It also involves rich human pose prior information. Particularly, for RGB stream 2D human pose can be acquired with the existing well-established 2D HPE approaches [43] pre-trained on large-scale RGB datasets (e.g., MS COCO [21]). Within depth stream, the physical 3D constraints on human body symmetry, skeleton length and joints' motion are maintained. Although the supervision priors above, to our knowledge, there is still no work that concerns applying unlabelled RGB-D pair sequence to address 3D HPE.

To fill this gap, a novel weakly supervised 3D HPE approach termed C3P for depth data is proposed by us, based on unlabelled RGB-D sequence. It *conducts cross-domain pose prior propagation from RGB to depth in weakly supervised manner, with self-supervised learning in depth domain jointly.* Human pose annotation supervision is only from the third-party 2D RGB datasets. To alleviate projection distortion [18], depth map will be transformed into point cloud.

For weakly supervised learning, the key idea is to build correspondence between 2D and 3D HPE results in 3D space. Compared with the existing 2D

supervision manner [7] for RGB-based 3D HPE, our method can better reveal 3D characteristics. Particularly after acquiring 2D HPE result on RGB stream with state-of-the-art 2D method [43], it will be back projected into 3D space in ray form according to RGB camera's intrinsic and extrinsic parameters. Then, for each predicted 3D joint its distance to the corresponding 2D oriented projection ray is minimized to establish accurate 3D to 2D correspondence as supervision.

To leverage performance, self-supervised learning in 3D domain is jointly executed. The supervision information derives from the intrinsic natural constraints on human body symmetry, skeleton length limitation and joint's temporal motion continuity thanks to cloud sequence's spatial-temporal characteristics.

Technically, the encoder-decoder based point set network (i.e., P2P [9]) for 3D hand pose estimation is used as our backbone network. Since under weakly-supervised setting joint's ground-truth 3D heatmap cannot be acquired for P2P's training, we propose to refine it with an additional per point aggregation module with integral regression design [37]. Accordingly, 3D heatmap is no longer required. Overall, our main research idea is shown in Fig. 1.

The experiments on CMU Panoptic [16, 17] and ITOP [11] datasets verify the effectiveness of our proposition. It is impressive that, when large-scale unlabelled data is introduced our weakly supervised approach can even outperform the fully supervised counterparts under the challenging cross-setup test setting.

The main contributions of this paper include:

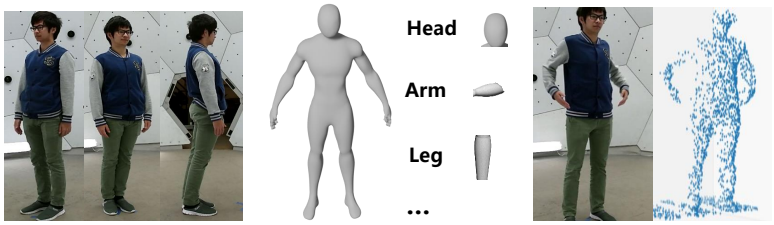
• We first propose the research problem of weakly supervised 3D human pose estimation in point cloud without requiring 3D annotation;

• C3P: a novel weakly supervised 3D human pose estimation approach that relies on unlabelled RGB-point cloud sequence.

2 Related work

In this section, we mainly introduce the depth image and point cloud based HPE methods. Since the proposed C3P is a weakly-supervised method, we also introduce other related weakly supervised HPE works.

Depth image and point cloud based HPE methods. Non-deep learning approaches [13, 34, 45] plays an important role in early research on depth image based human pose estimation. These methods rely on hand-crafted features and subsequent processing by regression or classification to obtain results on human posture. However, due to the limited discriminative power of these features, their performance is usually not as high as that of deep learning methods. Recently, numerous deep learning based HPE methods [27, 44, 52, 46, 48] were proposed due to the fitting ability of neural networks. V2V-PoseNet [27] converts depth image to Voxel, and uses 3D CNN to predict the coordinate of human joints, however, the 3D CNN is time-consuming. A2J [44] proposes an anchor based 2D CNN method which predicts in-plain offset estimation branch, depth estimation branch and anchor proposal branch, and uses the element-wise multiplication to get final results. A2J does not fully consider the intrinsic 3D information in



(a) Multi-view images. (b) Synthetic model. (c) RGB-D pair (ours).

Fig. 2. Illustration of the essential characteristics of different weakly-supervised 3D human pose estimation methods.

depth data, and the data augmentation strategies such as normalization, scale scaling will destroy the original 3D structure. Ge et al. [8, 9] propose a series of point cloud based 3D hand pose estimation methods. Ying et al. [46] propose an RGB-D based 3D HPE method, which extracts the joint point heatmap based on the RGB image, and then regresses the joint point offset and distance on the point cloud to obtain the final prediction results. Compared to Ying's method [46], our method only requires point clouds during testing. Note that the methods mentioned above are based on fully supervised training paradigm with labeled 3D data. However, collecting the annotated 3D dataset is labor and time consuming, leading to the fact that existing annotated 3D HPE datasets in depth form are generally of relative small size [11].

Weakly supervised HPE methods. The difficulty of data collection for 3D HPE leads researchers to use unlabeled data for facilitating model training. Here, the mainstream weakly supervised methods (i.e., multi-view and synthetic model based) are introduced as shown in Fig. 2. Most multi-view methods [19, 26, 15, 32] need images from three or more views and the complex calibration using intrinsic and extrinsic camera parameters among these views. Thus, accumulated errors are often faced. Towards this problem, some works only use two views [32] or try to predict camera parameters by deep network [40], which reduces the dependence of multi-view data. However compared with single-view data, collecting multi-view images is essentially difficult, which hinders the practical application. Another research avenue resorts to generating more training samples with synthetic software [35] or rendering predicted keypoints to depth or RGB images with pre-defined human skeleton model [39, 3, 2, 20]. Supervisory signals of synthetic data or rendered images facilitate training procedure of 3D pose estimator. However, these methods suffer from the domain shift problem between synthetic (or rendered) and real data. Human pose prior is also important for HPE. Some works use 2D annotation and limb proportions of human bodies [51, 49]. These methods often suffer from incomplete 3D information in RGB domain [51]. The prior information cannot be fully utilized, resulting in the need on fully annotated samples [51, 49]. In our work, unannotated RGB-point cloud pairs is used to link 2D and 3D domains. The 2D keypoints are obtained via pre-trained RGB pose estimator, which provides weak supervisory signals

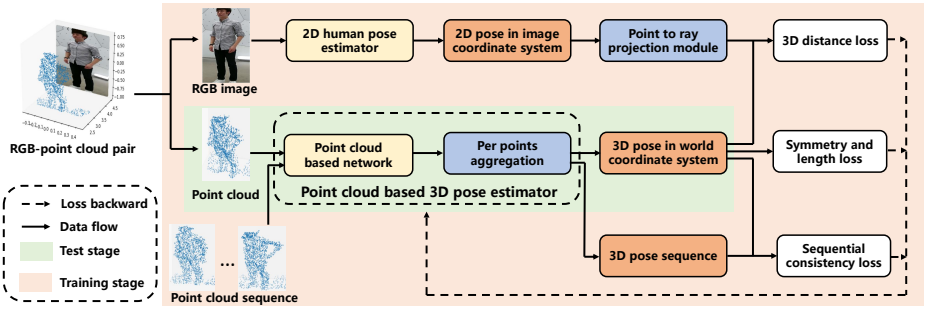


Fig. 3. The main technical pipeline of the proposed C3P approach.

for 3D HPE. To the best of our knowledge, there is still no work that concerns applying unlabelled RGB-D pair sequence to address 3D HPE problem.

3 Method

Here, C3P will be illustrated in details. It takes point cloud $P = \{p_i\}_{i=1}^{N}$ as input, and yields body joints' position $J = \{j_k\}_{k=1}^{K}$, where N is the number of points as input, K indicates the number of body joints, and $j_k \in R^3$ is the coordinate of body joints in real world distance w.r.t. camera. Each input point is of form $p_i \in R^{3+D}$, where the first 3 dimensions denote the coordinate in world coordinate system and $D = 3$ is the surface normal of points as in P2P [9].

C3P's main technical pipeline is shown in Fig. 3. In training stage, the input is RGB-point cloud sequence. Two point cloud networks (P2P [9] and PT [50]) are used as backbone to extract high level semantic features of raw points. Offset between each point and joint location is predicted. Under weakly-supervised setting, joint's ground-truth 3D heatmap cannot be obtained for training P2P or PT. Thus we propose to refine it with additional per point aggregation module with integral regression design [37], which integrally ensembles all points' prediction to obtain 3D joint location. In C3P, RGB based 2D pose estimator plays the role of providing weak supervision for training 3D pose estimator. A state-of-the-art 2D HPE method [43] pre-trained on MS COCO is adopted in C3P to predict accurate 2D joint location. Then, it will be back projected into 3D space in ray form according to RGB camera's intrinsic and extrinsic parameters. For each predicted 3D joint, its distance to the corresponding 2D oriented projection ray is minimized to establish accurate 3D to 2D correspondence as the supervision signal. Additionally, self-supervised learning in 3D domain is jointly proposed to leverage performance, including constraints on human body symmetry, skeleton length limitation and joint's temporal motion continuity.

In test stage, RGB image is no longer required. The weakly supervised P2P or PT will take point cloud as input and yield 3D joint position in world coordinate. Next, we will illustrate the weakly-supervised training part within C3P.

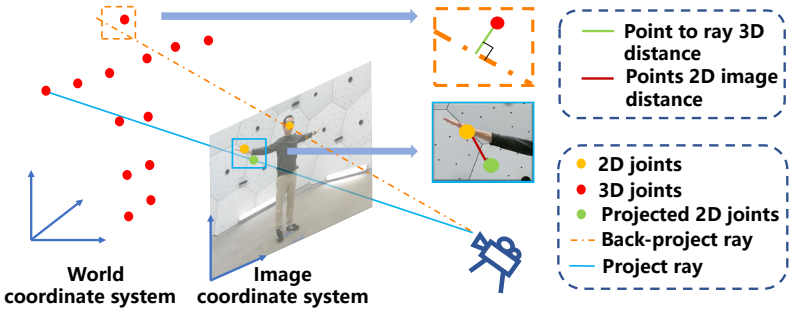


Fig. 4. Different supervision signals from RGB data. The solid line and sky blue box indicate that 3D keypoints are projected onto the image plane to calculate the planar pixel error. The dashed line and orange box indicate that the 2D keypoints are back-projected as 3D rays and the point-to-line distance error is calculated in 3D space.

3.1 Weak supervision signal from RGB image

As aforementioned, a 2D pose estimator [43] pre-trained on COCO dataset is used to get the RGB image based prediction results $C = \{c_k\}_{k=1}^{K}$, where K is the number of body joints, and $c_k \in R^2$ is the k-th joint on image plane. The point cloud based 3D pose estimator's output is $J = \{j_k\}_{k=1}^{K}$, and $j_k \in R^3$ is the k-th joint coordinate in world coordinate. Now the problem is how to use 2D joint prediction $C \in R^{2*K}$ to supervise the 3D output $J \in R^{3*K}$.

First, J is transformed into camera coordinate system using RGB camera's intrinsic and extrinsic parameters. We use $\hat{J} = \{\hat{j}_k\}_{k=1}^{K}$ to represent the 3D points in RGB camera coordinate system, and $\hat{C} = \{\hat{c}_k\}_{k=1}^{K}$ is the 2D projection results of $\hat{J}$. For one point $j_k = (x_k, y_k, z_k)^T$, $\hat{j}_k = (\hat{x}_k, \hat{y}_k, \hat{z}_k)^T$ and the projection result $\hat{c}_k = (\hat{u}_k, \hat{v}_k)^T$ is formulated as:

$$\left[\hat{x}_k\ \hat{y}_k\ \hat{z}_k\right]^T = R_{p-r}\left[x_k\ y_k\ z_k\right]^T + T_{p-r}, \tag{1}$$

$$\left[\hat{u}_k\ \hat{v}_k\ 1\right]^T = \frac{1}{\hat{z}_k} K_{rgb}\left[\hat{x}_k\ \hat{y}_k\ \hat{z}_k\right]^T, \tag{2}$$

where R_{p-r} and T_{p-r} are the rotation and translation matrix from point cloud to RGB camera coordinate system, K_{rgb} is RGB camera's intrinsic parameter.

One simple idea is to directly project the 3D joint predictions to 2D RGB image plane, distance between the projected joint locations $\hat{c}_k$ and 2D pose c_k can be utilized to formulate the loss function as:

$$\mathcal{L}_{rgb} = \sum_{k=1}^{K} \left\| c_k - \hat{c}_k \right\|_2 . \tag{3}$$

We argue that the above approach [7] may not be optimal. First, the optimization process of 3D estimator is different from 2D counterpart. That is, computing loss in 2D space inevitably yields discrepancy for training 3D network. Secondly

for weakly-supervised task, in the initial stage of training network's prediction results are indeed inaccurate. The projection process will result in huge loss in 2D plane when z is small, which leads to unstable training. And the unit of $\mathcal{L}_{rgb}$ is pixel, which may cause difficulty in 3D model tuning.

Accordingly, we propose to back-project the pre-computed keypoints c_k on image plane to ray forms in 3D space, and then calculate the distance between predicted 3D points and back-projected 3D rays as in Fig 4. The proposed back-projection method can yield more stable 3D training supervision signal.

Technically, the distance between j_k and the ray back-projected by c_k needs to be calculated. And, the point cloud and RGB camera coordinate system need to stay consistent. $\hat{j}_k$ is used to calculate divergence. $\mathcal{D}_{p-ray,k}$ is applied to denote distance between $\hat{j}_k = (\hat{x}_k, \hat{y}_k, \hat{z}_k)$ and the ray back-projected by $c_k = (u_k, v_k)$:

$$\mathcal{D}_{p-ray,k} = \frac{\left\| K_{rgb}^{-1} \hat{z}_k \begin{bmatrix} u_k & v_k & 1 \end{bmatrix}^T - \hat{j}_k \right\|_2}{\left\| K_{rgb}^{-1} \begin{bmatrix} u_k & v_k & 1 \end{bmatrix}^T \right\|_2}. \tag{4}$$

The weakly supervised signal from 2D keypoints is formulated as:

$$\mathcal{L}_{2d} = \sum_{k=1}^{K} \frac{\mathcal{D}_{p-ray,k}}{\mu}, \tag{5}$$

where we set the hyper parameter $\mu = \left\| \hat{j}_k \right\|_2$ to prevent the network from converging to trivial solution (i.e., zero vector).

3.2 Self supervision signal from point cloud

To further exploit the intrinsic prior information of human pose in depth cloud, we propose to build self-training supervision signals with the constraints on human body symmetry, bone length and joint's temporal motion continuity.

Bone length is an important prior knowledge of human bodies, especially in point cloud based method. We can acquire the real world scale, and estimate the absolute coordinates under world coordinate. This prior naturally provides useful information to regularize the output of 3D joint prediction as:

$$\mathcal{L}_{len} = \sum_{n=1}^{N} \left\| B_n - \bar{B}_n \right\|_2^2, \tag{6}$$

where N is the number of pre-defined bones; B_n is the n-th bone length, $\bar{B}_n$ is the mean length of the n-th bone across the dataset. With $\mathcal{L}_{len}$, the predicted 3D human poses are limited to a reasonable scale.

Human body symmetry is also an important prior. Being different from most existing RGB based methods that adopt 2.5D pose representation [14] (i.e, 2D image coordinates and depth related to root joint), C3P is under world coordinates. This makes the human symmetry prior easy to incorporate into

our proposition. Generally, the left bone (B_n) has almost equal length with the corresponding right counterpart (B_n^{cor}) as:

$$\mathcal{L}_{sym} = \sum_{n=1}^{N} \left\| \frac{B_n}{B_n^{cor}} - 1 \right\|. \tag{7}$$

It is worthy noting that, we do not use the form of loss like $\|B_n - B_n^{cor}\|$ since it may cause the final bone length approach to 0.

Motion consistency of joints in point cloud sequence is also a useful supervision signal. Intuitively, the movement of human joints in a temporal sequence is generally a continuous process. Requiring the 3D joints prediction to be continuous in the adjacent frames can well avoid jittering of ambiguous joints. In C3P, we concern the keypoints are moving at a constant speed:

$$\mathcal{L}_{con1} = \sum_{k=1}^{K} \sum_{i=2}^{I-1} \left\| j_k^i - \frac{j_k^{i-1} + j_k^{i+1}}{2} \right\|, \tag{8}$$

where I is number of video frames; j_k^i is location of k-th keypoint in i-th frame. In addition, the length of bones of the same person are constant in video:

$$\mathcal{L}_{con2} = \sum_{n=1}^{N} \left\| \frac{B_n}{\bar{B}_n^v} - 1 \right\|, \tag{9}$$

where $\bar{B}_n^v$ is the mean bone length in video for n-th bone.

The overall consistency loss is formulated as:

$$\mathcal{L}_{con} = \mathcal{L}_{con1} + \lambda_0 \mathcal{L}_{con2}, \tag{10}$$

where λ_0 is weight factor to balance two loss terms.

3.3 Per points aggregation

The network used by us for point cloud based 3D human pose estimation is P2P [9] and PT [50]. The two networks are designed for 3D hand pose estimation and point cloud semantic segmentation, which directly takes the 3D point cloud as input and yields dense prediction of each point. While our network output the per point heatmap and unit vector field to joint on the point cloud (similar to P2P [9]). The original P2P [9] uses the ground-truth heatmap to supervise the training procedure. However in weakly-supervised setting, there is no available annotation to generate 3D ground-truth heatmap and unit vector.

Towards this problem, we propose to use a points aggregation module to regress final predictions. During C3P's forward process, we select the results of the nearest points within a certain range to acquire the final joint prediction via weighted regression. The heatmap and unit vector field predictions of selected points integrally contribute to the final joint locations in ensemble manner [37].

This facilitates the training of 3D encoder-decoder based network without the need of ground-truth heatmap annotations.

In the predicted heatmap, the value of heatmap reflects the distance between current point and keypoint to be predicted, and the unit vector fields specify the direction. Small value in heatmaps means long distance. If the distance is above the threshold, this point will be abandoned when calculating the final result. The unit vector field specifies the direction between current point and target keypoint. Then, the result predicted by current point can be calculated.

The settings in our network enforces the final prediction results are surrounded by a set of points in point cloud. It is useful for our C3P method. Compared with the direct regression network [8], the dense prediction and per points aggregation module can guarantee the output of network at a relatively reasonable initial value, especially in the initial phase of training. This makes the weakly supervised training phase more stable.

3.4 Learning procedure of C3P

In the C3P's training stage, RGB-point cloud pairs are used as input. The 2D RGB human pose estimator is a frozen model that provides human pose in 2D space as weak supervision signal. The parameters in 3D depth-based network are updated with the 2D supervision and 3D self-supervision signals as well:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{2d} + \lambda_2 \mathcal{L}_{len} + \lambda_3 \mathcal{L}_{sym} + \lambda_4 \mathcal{L}_{con} \tag{11}$$

where $\mathcal{L}_{2d}$, $\mathcal{L}_{len}$, $\mathcal{L}_{sym}$, $\mathcal{L}_{con}$ are defined by Eq. (5)-(10). We set $\lambda_1 = 10$, $\lambda_2 = 1$, $\lambda_3 = 0.002$, $\lambda_4 = 0.1$ to balance each loss term.

In test phase, RGB images are no longer required. Given the raw 3D point cloud input, the trained 3D human pose estimator can yield 3D joint locations directly. This essentially expands the application scenarios of C3P.

4 Experiment

4.1 Datasets and evaluation metrics

CMU Panoptic Dataset [16, 17] contains 480 VGA videos, 31 HD videos and 10 Kinect videos, including depth and RGB images in indoor scene. Calibration and synchronous data are also provided. The 3D human pose annotation for VGA and HD videos are provided at the same time. We use 3D human pose in HD videos, the calibration clue, and synchronous data to get 3D human pose labels for Kinect. Since we only focus on point cloud based 3D human pose estimation for single person, experiments are carried on the "range of motion" class Kinect data with 9 available video sequences. The 6th and 9th videos ("171204_pose6, 171206_pose3") are used for testing, and the remaining 7 videos are for training. The evaluation metrics is the **mean average precision (mAP)** with 10-cm rule [11]. The average 3D distance error [9] is also used as evaluation metric.

Table 1. Performance comparison on CMU Panoptic Dataset [16, 17]. C3P (P2P) and C3P (PT) indicate C3P with the different backbone networks (i.e., P2P [9] and PT [50]). The unit of test error is cm.

	Nose	Eyes	Ears	Shoulders	Elbows	Wrists	Hips	Knees	Ankles	mAP	Error
Fully-supervised methods											
HandPoint[8]	79.8	78.9	79.6	80.6	5.3	0.2	89.6	84.5	75.1	62.8	12.0
P2P[9]	98.1	98.1	98.2	96.9	95.0	89.8	94.5	94.1	93.7	95.2	4.1
PT[50]	**99.6**	**99.3**	**99.2**	**98.8**	**97.0**	**92.0**	**96.8**	**95.0**	**95.7**	**96.9**	**3.3**
Weakly-supervised methods											
C3P(P2P)	96.3	95.8	95.0	93.9	91.4	81.5	90.9	78.4	85.2	89.4	6.1
C3P(PT)	**99.1**	**98.6**	**95.9**	**95.4**	**94.4**	**85.4**	**93.1**	**91.1**	**94.0**	**93.8**	**5.3**

ITOP Dataset[11] is a widely used benchmark dataset in depth image based and point cloud based 3D human pose estimation. It contains 40K training and 10K testing depth images and point clouds data for each front-view and top-view track. This dataset contains 20 actors and 15 human body parts are labeled with 3D coordinates relative to the depth camera. In our experiment, we use the front-view data to evaluate the effectiveness of the method. The evaluation metrics is the **mean average precision (mAP)** with 10-cm rule [11].

NTU RGB+D Dataset [33] is a large-scale RGB-D action recognition dataset. It contains over 40 subjects and 60 actions. The actions can cover most daily behavior. There are 17 different scenes in it. The size and diversity are much larger than current human pose dataset. This dataset also contains 3D skeleton joint position labeled by Kinect V2 SDK. But the annotation accuracy is not satisfactory. Nevertheless, the weakly-supervised setting in C3P can well leverage this large-scale RGB-D dataset. We conduct the cross-dataset test to demonstrate that, C3P trained on large-scale unannotated dataset can even better adapt to scene variation than the fully-supervised counterparts.

4.2 Implementation details

C3P is implemented using PyTorch. The input point number is set to 2048, and point cloud normalization operations is the same as P2P [9]. Adam is used as optimizer. The learning rate is set to 0.0001 in all cases. C3P is trained for 140 epoch. The learning rate decay by 0.1 at 90-th and 110-th epoch.

4.3 Comparison with fully-supervised method

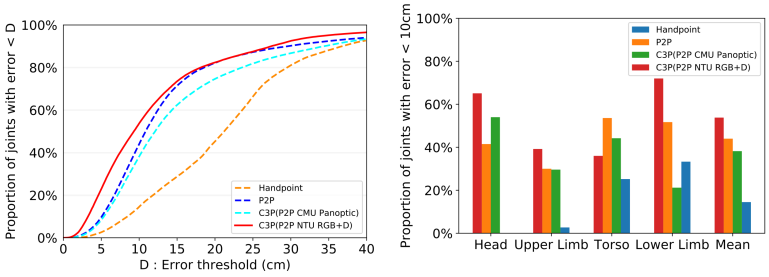
Here, C3P is compared with the state-of-the-art fully-supervised depth-based 3D HPE methods. The network structure and experimental setup are the same for the different datasets.

Results on CMU Panoptic dataset. C3P is compared with the fully-supervised 3D HPE methods (i.e., HandPoint [8], P2P [9], and PT [50]). Two different point cloud based networks (i.e., P2P[9] and PT[50]) are used to validate C3P's generality. Experimental results are listed in Table 1. We can see that:

• C3P achieves comparable results to fully-supervised counterparts. Without labeled data, its result is only inferior about 3% in mAP and 2cm in mean error;

Table 2. Performance comparison on ITOP Dataset [11]. C3P (P2P) and C3P (PT) indicate C3P with the different backbone networks (i.e., P2P [9] and PT [50]).

	Head	Neck	Shoulders	Elbows	Hands	Torso	Hips	Knees	Feet	mean
Fully-supervised methods										
RF[34]	63.8	86.4	83.3	73.2	51.3	65.0	50.8	65.7	61.3	65.8
RTW[47]	97.8	95.8	94.1	77.9	70.5	93.8	90.3	68.8	68.4	80.5
IEF[6]	96.2	85.2	77.2	45.4	30.9	84.7	83.5	81.8	80.9	71.0
VI[11]	98.1	97.5	96.5	73.3	68.7	85.6	72.0	69.0	60.8	77.4
CMB[41]	97.7	98.5	75.9	62.7	**84.4**	96.0	87.9	84.4	83.8	83.3
REN-9*6*6[10]	**98.7**	**99.4**	96.1	74.7	55.2	**98.7**	91.8	89.0	81.1	84.9
V2V*[27]	98.3	99.1	**99.2**	**80.4**	67.3	**98.7**	**93.2**	**91.8**	**87.6**	**88.7**
A2J[44]	98.5	99.2	96.2	78.9	68.4	98.5	90.9	90.8	86.9	88.0
P2P[9]	97.6	98.6	95.5	78.0	63.9	97.8	88.9	90.6	85.1	86.5
PT[50]	97.7	98.6	96.0	78.6	64.1	98.5	90.1	90.8	86.5	87.2
Weakly-supervised method										
C3P(P2P)	97.0	**97.4**	**91.4**	75.6	63.1	**96.4**	**85.2**	**88.7**	**84.2**	**84.5**
C3P(PT)	**97.3**	97.0	91.3	**76.3**	**66.5**	94.9	81.7	87.8	83.4	84.3



(a) Precision with different D (b) Joints' mean precision

Fig. 5. Results on cross view test setting. C3P (P2P CMU Panoptic) and C3P (P2P NTU RGB+D) indicate C3P models with P2P [9] and are trained on CMU panoptic [16, 17] and NTU RGB+D [33] respectively.

• C3P can be applied to the different point cloud networks [9, 50]. This indicates that the proposed weakly-supervised 3D HPE manner is not sensitive to the choice of backbone network;

• Although performance drop for joints (e.g., hand and feet) of fine appearance pattern, C3P's performance is still high (93.8 at mAP and 5.3mm at error).

Results on ITOP dataset. ITOP Dataset only contains depth images. Accordingly, we use the 2D annotation on depth image to replace the "supervision signal from RGB image" in C3P. State-of-art fully-supervised methods [44, 27, 10, 11, 41, 6, 47, 34] are compared to verify C2P's effectiveness. The performance comparison results are listed in Table 2. It can be observed that:

• C3P can achieve comparable results to fully-supervised methods. With the same 3D network, C3P's performance drop is slight (about 2-3% at mAP) .

• The results indicate that C3P can still work well when RGB images are missing. That is 2D annotation can also be used to train C3P.

Cross view test. To compare the generalization capacity of different 3D HPE methods, we conduct a cross view test on CMU Panoptic dataset. The

Table 3. Performance comparison between different weak supervision methods (i.e., our 2D-to-3D ray projection manner vs. 3D-to-2D projection way) on ITOP dataset.

	mAP									
	Head	Neck	Shoulders	Elbows	Hands	Torso	Hips	Knees	Feet	mean
3D-to-2D manner	95.9	95.0	90.0	67.8	48.6	91.7	80.5	85.6	76.5	78.7
2D-to-3D manner (ours)	**97.0**	**97.4**	**91.4**	**75.6**	**63.1**	**96.4**	**85.2**	**88.7**	**84.2**	**84.5**

Table 4. Effectiveness of self supervision signals within C3P.

Component	Mean error (cm)	mAP (@10cm)
w/o bone length $\mathcal{L}_{len}$	6.9	86.6
w/o human body symmetry $\mathcal{L}_{sym}$	6.6	87.5
w/o motion consistency $\mathcal{L}_{con}$	7.4	85.5
C3P (ours)	**6.1**	**89.4**

HPE methods are trained on one view and tested on another. To enhance generalization ability of deep learning model, previous efforts often resort to complex data augmentation strategies [9, 44]. However, C3P can use massive unannotated RGB-D data (e.g., NTU RGB+D [33] dataset) to achieve this goal. Specifically, we train C3P on NTU RGB+D dataset, and test it on CMU Panoptic dataset with cross view setting. The performance comparison among the different approaches is shown in Fig 5. We can summarize that:

• In this challenging test case, performance of all 3D HPE methods drops remarkably. However compared to the fully-supervised counterpart (i.e., P2P), C3P can consistently acquire better performance when large-scale unannotated dataset (i.e., NTU RGB+D) is used to enhance generality;

• The results above reveal that C3P can benefit from large-scale unannotated RGB-D data easy to collect, which is preferred by practical applications. C3P's performance tends to be further facilitated when more unannotated data is used.

4.4 Ablation study

Weak supervision signal. To verify the effectiveness of the proposed weak supervision information via 2D-to-3D ray projection against existing 3D-to-2D projection way, they are compared on ITOP dataset. 2D ground-truth human pose annotation is used to resist the effect of 2D HPE. The results are listed in Table 3. It can be observed that:

• The proposed 2D-to-3D ray projection is superior to existing 3D-to-2D counterpart. This is mainly due to the fact that, our method can better measure the distance between the predicted and target joints in world coordinate system;

• For joints (e.g., feet and hands) of high freedom degrees, the 3D-to-2D ray projection supervision outperforms the 3D-to-2D projection strategy by large margins (i.e., +7.7% on feet and +14.5% on hands) at mAP.

Self supervision signal. In our method, three self supervision signals (i.e, bone length, human body symmetry and temporal consistency constrains of human pose) are proposed. To validate their effectiveness, we conduct ablation

Table 5. Impact of 2D human pose estimator on CMU Panoptic dataset. C3P* indicates C3P model trained with 2D ground-truth pose. The unit of error is cm.

	Nose	Eyes	Ears	Shoulders	Elbows	Wrists	Hips	Knees	Ankles	mAP	error
C3P* (P2P)	98.1	98.2	97.6	95.8	93.4	82.8	94.5	82.3	91.4	92.4	4.7
C3P* (PT)	**99.2**	**98.9**	**98.4**	**97.1**	**96.2**	**89.5**	**96.0**	**94.8**	**95.3**	**96.0**	**3.7**
C3P (P2P)	96.3	95.8	95.0	93.9	91.4	81.5	90.9	78.4	85.2	89.4	6.1
C3P (PT)	**99.1**	**98.6**	**95.9**	**95.4**	**94.4**	**85.4**	**93.1**	**91.1**	**94.0**	**93.8**	**5.3**

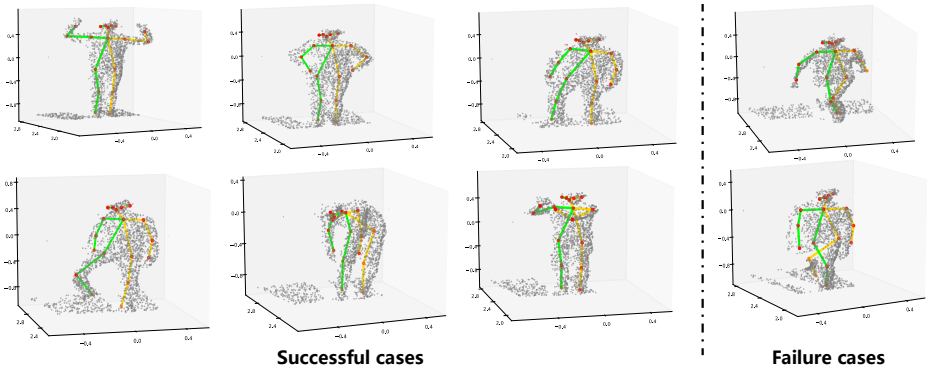


Fig. 6. C3P's qualitative results on CMU Panoptic dataset.

test on CMU Panoptic dataset by removing them item by item respectively. The results are listed in Table 4. It can be observed that:

• All the three self supervision signals are essential for leveraging performance. Among them, the temporal consistency constraint contributes the most towards the final result (i.e., 1.3cm at mean error and +3.9% at mAP).

Impact of 2D human pose estimator. 2D human pose estimator plays the important role for generating C3P's weak supervision signals. Hence the effectiveness of 2D pose estimator can affect C3P's training phase remarkably. To investigate the impact of 2D human pose estimator, we compare it with the ground-truth 2D annotation on CMU Panoptic dataset. The results are listed in Table 5. It can be observed that:

• The impact of 2D pose estimator is remarkable to C3P. More accurate 2D pose estimation result leads to better 3D HPE;

• The performance gain brought by more accurate 2D keypoint locations is consistent across all human joints. This reveals that, improving the quality of weak supervision signals (i.e., 2D pose estimation) is critical to C3P.

4.5 Qualitative results

Some C3P's qualitative results on CMU Panoptic Dataset and NTU RGB+D Dataset are shown in Fig. 6. Generally, C3P works well towards variational human poses. The failure cases are mainly due to serious self-occlusion and missing points in point cloud. While, we also find that C3P is still of some anti-occlusion

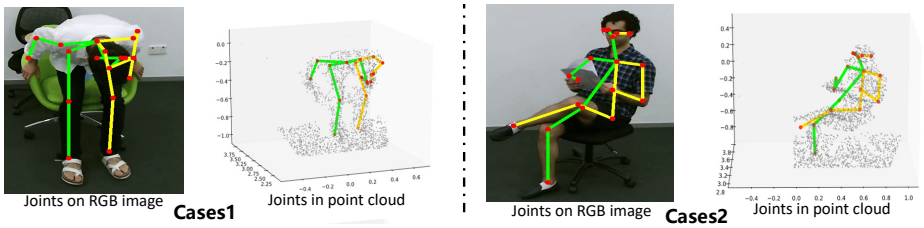


Fig. 7. C3P's qualitative results for anti-occlusion on NTU RGB+D dataset.

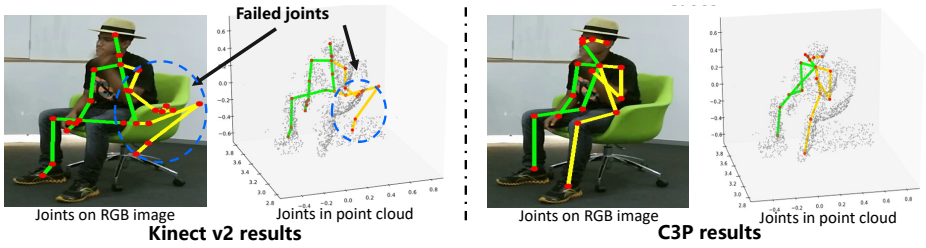


Fig. 8. Intuitive comparison between C3P and MicroSoft Kinect V2 SDK on NTU RGB+D dataset.

capacity as shown in Fig. 7, which may be due to the introduction of self supervision signals on temporal consistence and bone length. C3P is also compared with Kinect V2 SDK in Fig. 8. Under human-object interaction condition, C3P even outperforms Kinect V2. This indeed verifies C3P's application potentiality.

5 Conclusions

In this paper, C3P is proposed as a novel weakly supervised 3D HPE method towards point cloud. Its training phases does not require any 3D human pose annotation. Instead, we propose to propagate the 3D pose prior within the unlabelled RGB-point cloud sequence to 3D domain. The supervision signals derive from the well-established 2D pose estimator and the physical constrains of 3D human body. Extensive experiments demonstrate that C3P can achieve comparable or even better performance than the fully supervised counterparts. How to enhance C3P's anti-occlusion capacity is what we mainly concern in future.

Acknowledgements. This work is jointly supported by the National Natural Science Foundation of China (Grant No. 61502187 and 61876211). Joey Tianyi Zhou is supported by SERC Central Research Fund (Use-inspired Basic Research), Programmatic Grant No. A18A1b0045 from the Singapore government's Research, and Innovation and Enterprise 2020 plan (Advanced Manufacturing and Engineering domain).

References

1. Caetano, C., Sena, J., Brémond, F., Dos Santos, J.A., Schwartz, W.R.: Skelemotion: A new representation of skeleton joint sequences based on motion information for 3d action recognition. In: Proceedings of IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). pp. 1–8 (2019)
2. Cai, Y., Ge, L., Cai, J., Thalmann, N.M., Yuan, J.: 3d hand pose estimation using synthetic data and weakly labeled rgb images. IEEE transactions on pattern analysis and machine intelligence **43**(11), 3739–3753 (2020)
3. Cai, Y., Ge, L., Cai, J., Yuan, J.: Weakly-supervised 3d hand pose estimation from monocular rgb images. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 666–682 (2018)
4. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
5. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
6. Carreira, J., Agrawal, P., Fragkiadaki, K., Malik, J.: Human pose estimation with iterative error feedback. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4733–4742 (2016)
7. Chen, X., Lin, K.Y., Liu, W., Qian, C., Lin, L.: Weakly-supervised discovery of geometry-aware representation for 3d human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). pp. 10895–10904 (2019)
8. Ge, L., Cai, Y., Weng, J., Yuan, J.: Hand pointnet: 3d hand pose estimation using point sets. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8417–8426 (2018)
9. Ge, L., Ren, Z., Yuan, J.: Point-to-point regression pointnet for 3d hand pose estimation. In: Proceedings of the European conference on computer vision (ECCV). pp. 475–491 (2018)
10. Guo, H., Wang, G., Chen, X., Zhang, C.: Towards good practices for deep 3d hand pose estimation. arXiv preprint arXiv:1707.07248 (2017)
11. Haque, A., Peng, B., Luo, Z., Alahi, A., Yeung, S., Fei-Fei, L.: Towards viewpoint invariant 3d human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 160–177 (2016)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
13. He, L., Wang, G., Liao, Q., Xue, J.H.: Depth-images-based pose estimation using regression forests and graphical models. Neurocomputing **164**, 210–219 (2015)
14. Iqbal, U., Molchanov, P., Gall, T.B.J., Kautz, J.: Hand pose estimation via latent 2.5d heatmap regression. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
15. Iqbal, U., Molchanov, P., Kautz, J.: Weakly-supervised 3d human pose learning via multi-view images in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5243–5252 (2020)
16. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015)

17. Joo, H., Simon, T., Li, X., Liu, H., Tan, L., Gui, L., Banerjee, S., Godisart, T.S., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social interaction capture. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
18. Kim, W.S., Ortega, A., Lai, P., Tian, D., Gomila, C.: Depth map distortion analysis for view rendering and depth coding. In: Proceedings of the IEEE International Conference on Image Processing (ICIP). pp. 721–724 (2009)
19. Kocabas, M., Karagoz, S., Akbas, E.: Self-supervised learning of 3d human pose using multi-view geometry. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1077–1086 (2019)
20. Kundu, J.N., Seth, S., Jampani, V., Rakesh, M., Babu, R.V., Chakraborty, A.: Self-supervised 3d human pose estimation via part guided novel image synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6152–6162 (2020)
21. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proceedings of the European conference on computer vision (ECCV). pp. 740–755. Springer (2014)
22. Liu, J., Wang, G., Duan, L.Y., Abdiyeva, K., Kot, A.C.: Skeleton-based human action recognition with global context-aware attention lstm networks. IEEE Transactions on Image Processing **27**(4), 1586–1599 (2017)
23. Liu, J., Wang, G., Hu, P., Duan, L.Y., Kot, A.C.: Global context-aware attention lstm networks for 3d action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1647–1656 (2017)
24. Microsoft: Kinect for windows. `https://developer.microsoft.com/en-us/windows/kinect/`, accessed February 6, 2022
25. Microsoft: Kinect for x-box 360. `http://www.xbox.com/en-US/kinect`, accessed February 6, 2022
26. Mitra, R., Gundavarapu, N.B., Sharma, A., Jain, A.: Multiview-consistent semi-supervised learning for 3d human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6907–6916 (2020)
27. Moon, G., Chang, J.Y., Lee, K.M.: V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5079–5088 (2018)
28. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 483–499 (2016)
29. Obdržálek, Š., Kurillo, G., Han, J., Abresch, T., Bajcsy, R.: Real-time human pose detection and tracking for tele-rehabilitation in virtual reality. In: Medicine Meets Virtual Reality 19, pp. 320–324. IOS Press (2012)
30. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 652–660 (2017)
31. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017)
32. Remelli, E., Han, S., Honari, S., Fua, P., Wang, R.: Lightweight multi-view 3d pose estimation through camera-disentangled representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6040–6049 (2020)

33. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1010–1019 (2016)
34. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1297–1304 (2011)
35. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2107–2116 (2017)
36. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
37. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 529–545 (2018)
38. Svenstrup, M., Tranberg, S., Andersen, H.J., Bak, T.: Pose estimation and adaptive robot behaviour for human-robot interaction. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). pp. 3571–3576 (2009)
39. Wan, C., Probst, T., Gool, L.V., Yao, A.: Self-supervised 3d hand pose estimation through training by fitting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10853–10862 (2019)
40. Wandt, B., Rosenhahn, B.: Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7782–7791 (2019)
41. Wang, K., Lin, L., Ren, C., Zhang, W., Sun, W.: Convolutional memory blocks for depth data representation learning. In: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI). pp. 2790–2797 (2018)
42. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
43. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
44. Xiong, F., Zhang, B., Xiao, Y., Cao, Z., Yu, T., Zhou, J.T., Yuan, J.: A2j: Anchor-to-joint regression network for 3d articulated pose estimation from a single depth image. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 793–802 (2019)
45. Ye, M., Wang, X., Yang, R., Ren, L., Pollefeys, M.: Accurate 3d pose estimation from a single depth image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). pp. 731–738 (2011)
46. Ying, J., Zhao, X.: Rgb-d fusion for point-cloud-based 3d human pose estimation. In: Proceedings of IEEE International Conference on Image Processing (ICIP). pp. 3108–3112 (2021)
47. Yub Jung, H., Lee, S., Seok Heo, Y., Dong Yun, I.: Random tree walk toward instantaneous 3d human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2467–2474 (2015)
48. Zhang, B., Xiao, Y., Xiong, F., Wu, C., Cao, Z., Liu, P., Zhou, J.T.: 3d human pose estimation with cross-modality training and multi-scale local refinement. Applied Soft Computing **122**, 108950 (2022)

49. Zhang, Z., Hu, L., Deng, X., Xia, S.: Weakly supervised adversarial learning for 3d human pose estimation from point clouds. IEEE transactions on visualization and computer graphics **26**(5), 1851–1859 (2020)
50. Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V.: Point transformer. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 16259–16268 (2021)
51. Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y.: Towards 3d human pose estimation in the wild: a weakly-supervised approach. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 398–407 (2017)
52. Zhou, Y., Dong, H., El Saddik, A.: Learning to estimate 3d human pose from point cloud. IEEE Sensors Journal **20**(20), 12334–12342 (2020)