Weakly Supervised Adversarial Learning for 3D Human Pose Estimation from Point Clouds

Zihao Zhang, Lei Hu, Xiaoming Deng, and Shihong Xia

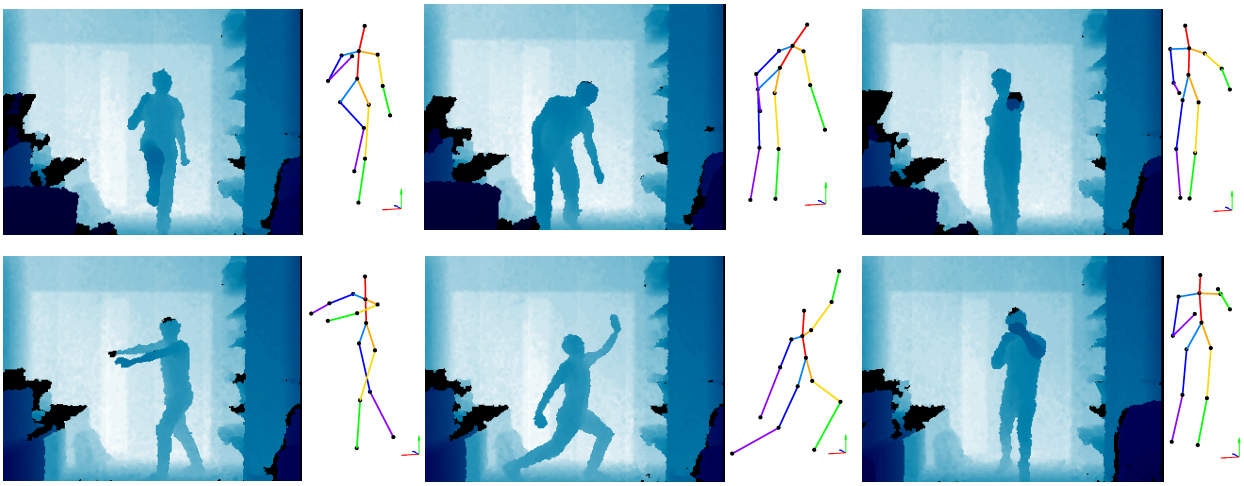


Fig. 1. Examples of qualitative results on ITOP dataset using our 3D human pose estimation method. (The estimated 3D human pose may be not shown under the camera viewpoint)

Abstract— Point clouds-based 3D human pose estimation that aims to recover the 3D locations of human skeleton joints plays an important role in many AR/VR applications. The success of existing methods is generally built upon large scale data annotated with 3D human joints. However, it is a labor-intensive and error-prone process to annotate 3D human joints from input depth images or point clouds, due to the self-occlusion between body parts as well as the tedious annotation process on 3D point clouds. Meanwhile, it is easier to construct human pose datasets with 2D human joint annotations on depth images. To address this problem, we present a weakly supervised adversarial learning framework for 3D human pose estimation from point clouds. Compared to existing 3D human pose estimation methods from depth images or point clouds, we exploit both the weakly supervised data with only annotations of 2D human joints and fully supervised data with annotations of 3D human joints. In order to relieve the human pose ambiguity due to weak supervision, we adopt adversarial learning to ensure the recovered human pose is valid. Instead of using either 2D or 3D representations of depth images in previous methods, we exploit both point clouds and the input depth image. We adopt 2D CNN to extract 2D human joints from the input depth image, 2D human joints aid us in obtaining the initial 3D human joints and selecting effective sampling points that could reduce the computation cost of 3D human pose regression using point clouds network. The used point clouds network can narrow down the domain gap between the network input i.e. point clouds and 3D joints. Thanks to weakly supervised adversarial learning framework, our method can achieve accurate 3D human pose from point clouds. Experiments on the ITOP dataset and EVAL dataset demonstrate that our method can achieve state-of-the-art performance efficiently.

Index Terms—Human Pose Estimation, Point Clouds, Depth Map

1 Introduction

Human pose estimation plays a key role in applications such as AR/VR, special effects and human-computer interactions, and it can help in the understanding of the intentions underlying interactions and provide proper feedback. Recently, with the boom in the development of depth sensors, rapid progress has been made to recover 3D human poses using point clouds as input, which has been one of the major image-based approaches to achieve high-quality 3D human poses.

Though many efforts on human pose estimation have been made in recent literature, point clouds based 3D human pose estimation is still challenging. First, occlusions due to multiple persons and human body self-occlusion can hinder the performance of human pose estimation. Secondly, it is a labor-intensive and error-prone process to manually label 3D human joints from input depth or point clouds, due to the self-occlusion between body parts as well as tedious annotation on 3D point clouds. Thirdly, it is a key issue to represent or sample the input point clouds in an effective manner. The methods [12, 25] project 3D point clouds to 2D representation and use 2D convolution neural network, which could lose the 3D context of the point clouds. Due to the domain gap between 2D depth and 3D human joints, it is challenging to learn the mapping between 2D depth and 3D human joints. Several methods adopt 3D volume (voxel or TSDF) [19, 33, 39] or point clouds representations [24, 26]. For 3D volume representation, 3D CNN is often used for feature extraction and can achieve high-quality human pose performance. However, the network contains more network parame-

- *Zihao Zhang is with Institute of Computing Technology, Chinese Academy of Sciences and the University of Chinese Academy of Sciences. E-mail: zhangzihao@ict.ac.cn.*
- *Lei Hu is with Institute of Computing Technology, Chinese Academy of Sciences and the University of Chinese Academy of Sciences. He share the same contribution as Zihao Zhang. E-mail: hulei19z@ict.ac.cn.*
- *Xiaoming Deng is with Bejing Key Laboratory of Human Computer Interactions, Institute of Software, Chinese Academy of Sciences. E-mail: xiaoming@iscas.ac.cn.*
- *Shihong Xia is with Institute of Computing Technology, Chinese Academy of Sciences and the University of Chinese Academy of Sciences. He is the corresponding author. E-mail: xsh@ict.ac.cn.*

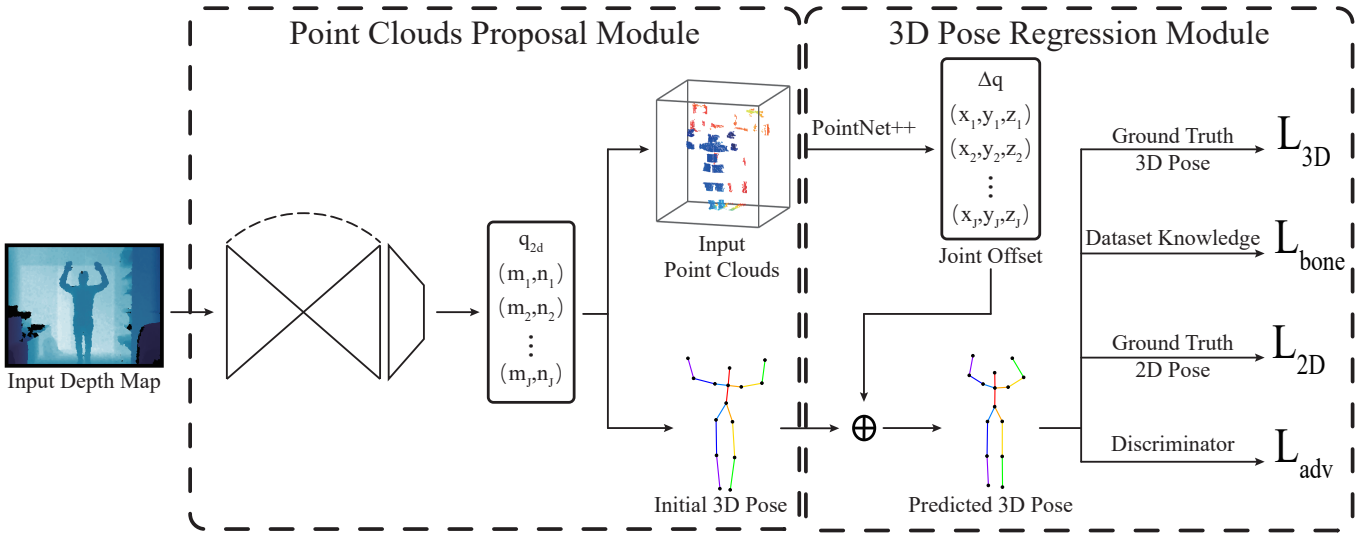


Fig. 2. Overview of our 3D human pose estimation network. The network consists of two modules, the point clouds proposal module and the 3D pose regression module. Using the input depth map, we first estimate the 2D human pose, and use it to sample and normalize the extracted point clouds from depth. Then we use the initial 3D pose converted from the estimated 2D pose and the normalized point clouds to predict the final 3D human pose.

ters that those networks using depth as input and is computationally expensive. For point clouds representation, existing PointNet-based methods [24, 26] have the advantage of light-weighted networks, and often adopt point sampling strategy to reduce the computational effort due to the massive input points, while it may affect local details of the input and decrease the pose estimation performance. It is a key issue to design an effective point sampling strategy.

To address the difficulty of 3D human joint annotation, we observe that it is easier to construct human pose datasets with 2D human joint annotations on depth image. For example, we can label 2D human joints on depth image via aligned color image using an off-the-shelf RGBD camera. Though 2D human joints is closely related to 3D human pose, it can only provide weak supervision of 3D human pose. Therefore, it is challenging to design an effective approach to exploit fully labeled dataset with 3D joints as well as weakly labeled data with 2D joints only to achieve better 3D human pose estimation performance.

In this paper, we propose a new human pose estimation network from point clouds, which can be trained in a weakly-supervised manner. The key idea of our method is that, the network is designed to predict 3D human poses so that during the training stage for fully labeled data the estimated 3D joints are well matched with the ground truth 3D joints, and for weakly labeled data that contain annotations of 2D joints the 2D projection of estimated 3D joints can align well with the ground truth 2D joints. To handle the possible ambiguity from the weakly labeled data, we adopt a discriminator network to judge whether the reconstructed human joints are plausible or not and enforce the bone length ratios. Our method consists of two modules, point clouds proposal module and 3D pose regression module. In the point clouds proposal module, we extract 2D human pose via a 2D heat map from the input depth using a compact yet effective fully convolutional network, and use the 2D human pose to sample the point clouds. Then the sampled point cloud is normalized with the estimated root joint; In the 3D pose regression module, we recover 3D human pose following a generative adversarial network (GAN). For the generator, we recover the human pose using the hierarchical PointNet-based regression with the sampled point clouds as input. For the discriminator, we use a fully connected neural network to distinguish the estimated human pose from the ground truth human pose. Experimental results show that our method achieves state-of-the-art results, and that weakly supervised learning with additional images with 2D joints is effective. Our method can directly estimate the human pose from the point clouds with background information, and we demonstrate the reliability of our approach on the ITOP dataset [12] and EVAL dataset [8].

Compared to existing methods using either 2D or 3D representations of depth images, we exploit both point clouds and the input depth images. We adopt 2D CNN to extract 2D human joints from depth images, 2D human joints could help us get initial 3D human joints and select sample points that reduce the computation cost of the 3D pose regression network built upon PointNet and enhance the performance of the human pose estimation.

The main contributions of our work can be summarized as follows.

1 We propose a new weakly-supervised deep learning network for the 3D human pose estimation problem. We exploit both the weakly supervised data with only annotations of 2D human joints and fully supervised data with annotations of 3D human joints. To relieve the human pose ambiguity due to weak supervision, we adopt adversarial learning to ensure that the recovered human pose is valid. Experimental results demonstrate that with the additional weakly supervised data, our network could achieve better results than the network with only fully supervised data. Our research can inspire human pose/shape recovery tasks in which full supervision, such as 3D joints or shapes, is not available.

2 Compared to existing human pose estimation methods that require human foreground detection/segmentation, our method is capable of human pose estimation without explicit human foreground detection/segmentation.

3 Our method achieves state-of-the-art, time-efficient 3D human pose estimation performance on the ITOP and EVAL datasets.

2 Related Works

In this section, we review related researches including 3D human pose estimation, weakly supervised learning, and neural networks for 3D representations.

2.1 3D Human Pose Estimation

3D human pose estimation methods can be divided into two categories, namely, the discriminative methods and the generative methods.

Discriminative Methods. Discriminative methods directly estimate 3D human pose from the input image. Present random forest based methods can be classified into discriminative category [10, 28, 35]. Recently, several convolutional neural network (CNN) based methods

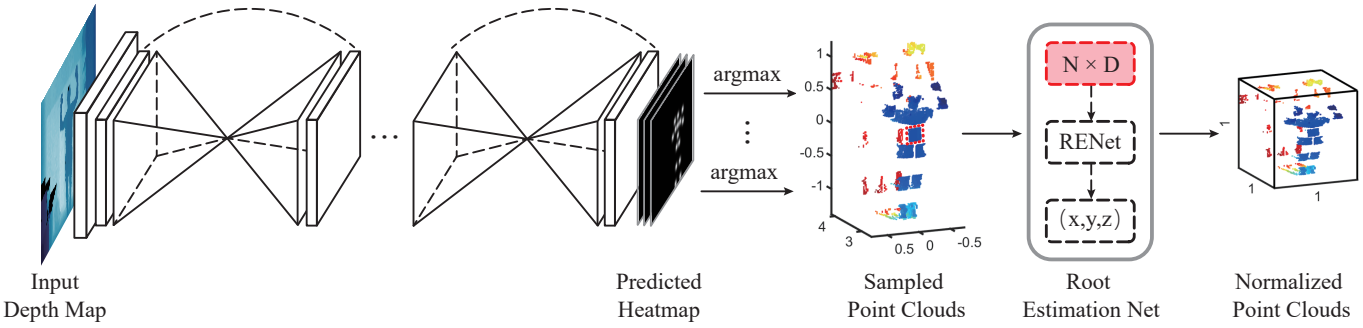


Fig. 3. Illustration of our point clouds proposal module. We detect 2D joints via joint heatmaps, use the 2D joints to sample the point clouds, and the sampled point clouds is then normalized by the estimated position of root joint and the size of a pre-defined bounding box.

are used to predict 3D human pose under the discriminative framework. Pavlakos *et al.* [21] use a 3D volumetric representation for 3D human pose, and adopt CNN to predict the likelihood of 3D joints in each voxel. Sun *et al.* [30] propose an integral operation which combines the task of heatmap representation and joint regression. In this way, they could avoid the non-differentiable post-processing and quantization error caused by heatmap representation. Marin-Jimenez *et al.* [16] design a representation for 3D human pose via a linear combination of the pose prototypes, and then use that representation to estimate the 3D human pose from depth. Though discriminative methods can directly get 3D human pose from the image, these methods are severely affected by the scarcity of training data and could achieve poor accuracy of human pose estimation for extraordinary poses.

Generative Methods. Generative methods often use the 2D information extracted from an input image to infer reasonable 3D poses for the image. These methods have the advantage that the 2D information extraction step only requires data with easily-accessed annotations such as 2D joint locations [2, 3, 5, 6, 29, 31, 34, 36]. Martinez *et al.* [17] use a very simple yet effective combination of linear layer to lift the 2D human pose to 3D human pose, and can achieve reasonable results. In 3D human pose estimation with depth images, Wang *et al.* [32] use a fully convolutional network to estimate the 2D pose in a depth image, and predict the 3D pose by an inference built-in MatchNet [11]. Haque *et al.* [12] present a heatmap-like 2D glimpse from a depth image, and use the recurrent network to predict the 3D pose in an iterative manner. However, the performance of generative 3D human pose methods is largely affected by the accuracy of 2D information extraction.

2.2 Weakly-supervised Learning

In recent years, we witnessed the fast development of deep learning methods. However, most of the deep learning methods require large-scale labeled training data. It is well established that annotating large-scale data is expensive and time-consuming. For 3D human pose estimation using point clouds, we also suffer from the lack of large-scale training data due to the fact that it is even more difficult to annotate 3D human joints from point clouds.

Followed by the guideline in [40], we divide the weak supervision learning into three types. The first type [4,38] is *incomplete* supervision, where the training data are a mixture of unlabeled data and a small amount of labeled data. In this case, the labeled data is too limited to train a good model, while the unlabeled or semi-labeled data are very easy to acquire. The second type [7, 37] is *inexact* supervision, where the labels are coarse or inexact. The third type [1, 9] is *inaccurate* supervision, which means the given label may not be the ground truth.

As for 3D human pose estimation scenario, we mainly discuss the *incomplete* supervision about the weakly-supervised learning methods. Recently, several 3D human pose estimation methods using weakly-supervised manner are proposed [14, 22, 27, 38]. Zhou *et al.* [38] propose a weakly supervised method for color-based 3D human pose estimation to use the 2D joint labels and 3D joint labels to predict 3D human pose. Pavlakos *et al.* [22] propose a pictorial structure model to predict human pose from the multi-view heatmaps. In [27], the authors take the multi-view consistency as constraint, which makes them only need a few data with 3D label. Kocabas *et al.* [14] propose to use paired multi-view images to predict the 3D human pose under the constraint of epipolar geometry. These methods are all designed for 3D human pose estimation using color images as input. So far, less attentions are paid on weakly-supervised learning of 3D human pose using depth image or point clouds as input.

2.3 Neural Networks for 3D Representations

3D representations such as point clouds and 3D volume are effective for scene understanding and pose estimation, and for these representations, point clouds networks and voxel-based CNNs are widely used models.

Point Clouds Networks. The point clouds networks directly use point clouds as input. Qi *et al.* [24] propose an end-to-end network named PointNet, which extracts point-wise features and can be applied to 3D object classification and point-level semantic segmentation tasks. Qi *et al.* [26] propose PointNet++, which enhances PointNet by learning the local structure on different scales. Motivated by the idea of using convolution on point clouds, Li *et al.* [15] propose PointCNN that predicts a transformation matrix and applies it before the convolution, and achieve the state-of-the-art performance on 3D object classification and point-level semantic segmentation tasks. To accelerate the 3D detection in hybrid camera, Qi *et al.* [23] propose the Frustum PointNet, which uses the 2D information to get a frustum and use the PointNet to get the 3D object detections. This method is more time-efficient due to the fact that 2D detection can reduce the point clouds region that requires object detection effort.

Voxel-based CNNs. Voxel-based CNN methods usually convert the point clouds into 3D volumetric representation and use 3D CNN to extract features [18, 25]. Zhou *et al.* [39] propose an end-to-end framework to predict the object's bounding box in the 3D space by dividing the 3D point clouds into 3D voxels. In 3D human pose estimation, Chang *et al.* [19] voxelize the point clouds and apply 3D CNN to the voxelized point clouds to predict the 3D pose.

3 Methodology

We propose a 3D human pose estimation architecture from a single depth map under a GAN framework, which reconstructs the locations of 3D human joints $q \in \mathscr{R}^{J\times 3}$ ($\mathscr{R}$ is the set of real numbers, J is the joint number) in the depth camera coordinate system. As shown in Fig. 2, our generator network aims to reconstruct human pose and consists of two modules: a point cloud proposal module and a 3D pose regression module. The generated human joints from the last stage are judged by the discriminator network to be real or fake. The details of the generator network and discriminator network will be elaborated in the following sections.

3.1 Preliminary: PointNet and PointNet++

We first provide some background about two representative point clouds-based networks; a more detailed introduction can be found

in [24, 26]. PointNet is a network architecture that can extract features from 3D unordered point clouds. PointNet usually takes the point coordinate and other features such as surface normals as input, which are then mapped to a higher-dimensional space using a multi-layer perceptron. The main limitation of PointNet is its inability to capture the local structure of the point clouds metric space, which makes it difficult to understand the detailed spatial pattern. To solve this problem, Qi *et al.* [26] proposed a hierarchical PointNet named PointNet++, which defines the partitioning of point clouds using the neighborhood in Euclidean space (i.e. the sampling and grouping step of PointNet++) and applies PointNet recursively to the neighborhood. Given the point clouds $p \in R^{M,c}$, where M is the point number and c is the feature dimension of the point clouds, PointNet++ adopts farthest point sampling to select M_1 points as the centroid of the neighborhood, and then uses PointNet to extract the features from the k-nearest neighbor. The extracted features are then recursively fed into the next sampling level until the level reaches the manually defined level.

3.2 Point Clouds Proposal Module

The point clouds proposal module aims to design an effective point clouds sampling via 2D human pose, which could improve the efficiency in the 3D pose regression module (See Section 3.3). The module consists of two steps, i.e. 2D pose detection, and point clouds sampling and normalization.

2D Pose Detection. We perform 2D human pose estimation according to the procedure of Newell *et al.* [20] due to the compactness and high performance of their network. Details of the network structure are shown in the left part of Fig. 3.

We set the loss function as the L_2 distance between the predicted heat map and the heat map generated from the ground truth 2D joint locations q^*_{2d}. The 2D human pose $q_{2d} \in \mathscr{R}^{J\times 2}$ can be recovered by arg max operation on the predicted heat map. Notice that arg max operation is not continuous or differentiable, we have to train this network separately. We pretrain this model on our synthesis dataset and fine-tune it on the target dataset. The average 2D joint error with our trained model on the ITOP dataset is below 5 pixels.

Point Clouds Sampling and Normalization. The above 2D joint detection can guide us to get pose-aware sampled point clouds to recover the 3D human pose. We crop the bounding box of detected 2D joints from depth map, then extract J local image patches of $d = 20\times 20$ pixels centered at the detected 2D joints. With the intrinsic matrix of the depth camera, we backproject the image patches to 3D point clouds $p \in \mathscr{R}^{N\times 3}$ ($N = d\times J$ is the number of the total sampled points), named the sampled point clouds hereafter.

Then we get the position of root joint p_{root} by feeding the sampled point clouds to a tiny regression network with ResNet backbone, named root estimation network (RENet). With the root joint p_{root}, the sampled point clouds can be normalized to $[-1,1]^3$ by:

$$p_{norm} = \frac{p - p_{root}}{L} \tag{1}$$

where p_{norm} is the normalized point clouds of p, p_{root} is the position of estimated root joint, and L is the size of a predefined bounding box, set to $L = [1.5, 1.5, 2]$ in our experiments. We apply this transformation to the sampled point clouds, the ground truth and the initial 3D human pose.

3.3 3D Pose Regression Module

The 3D pose regress module aims to predict the 3D human pose via the 2D human pose and the sampled point clouds. We first obtain the initial pose q_{init} by backprojecting the estimated 2D pose q_{2d} into 3D space and setting the depth of each joint to our estimated depth of the root joint approximately. Then we estimate the joint offsets Δq to the initial pose q_{init}, and the estimated pose q can be updated by $q = q_{init} + \Delta q$.

3.3.1 Network Architecture

Our 3D pose regression network follows GAN architectures, consists of generator and discriminator networks.

Generator The generator aims to obtain the joint offsets Δq of q_{init}. The network architecture of the generator i.e. the 3D pose regression module is shown in Fig. 2. The inputs of the 3D pose regression network are the normalized point clouds p_{norm}. We feed p_{norm} to hierarchical PointNet [26], and obtain the joint offsets Δq. The hierarchical PointNet uses has three point-abstraction levels. The number of local regions in each level is $N_1 = 64, N_2 = 32, N_3 = 8$, and each local region contains $k = 8$ points. The extracted features of each level are $C_1 = 128, C_2 = 256, C_3 = 1024$, respectively. After we obtain the joint offsets Δq, we can update the 3D human pose by $q = q_{init} + \Delta q$.

Discriminator The discriminator aims to judge whether the recovered 3D human pose is real or fake. Fig. 4 shows the structure of our discriminator network. The network uses the predicted 3D joint position as input, and predicts whether the predicted 3D joint position is real or fake. The discriminator consists of two fully connected layers (FC) with leaky RELU functions and skip connections, one fully connected layer with a leaky ReLU function, another fully connected layer, and a softmax function.

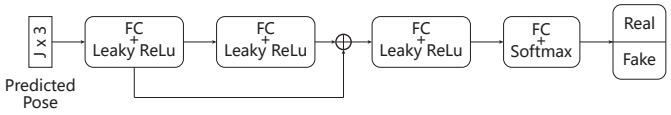


Fig. 4. The network structure of the discriminator.

3.3.2 Loss Function

We design a 3D pose regression network using both the fully labeled data and the weakly labeled data. For the fully labeled data, i.e. $\mathscr{S}_{full} = \{I, q^*_{2D}, q^*_{3D}\}$, we use full 3D joint supervision to enforce 3D human joint supervision, and we adopt 2D joint projection loss, 3D joint loss and bone length ratio loss. For the weakly labeled data, i.e. $\mathscr{S}_{weak} = \{I, q^*_{2D}\}$, we can only adopt weak supervision of 3D joints with the 2D joint annotations, and use the 2D joint projection loss and bone length ratio loss.

The loss function of our generator can be defined as follows:

$$\begin{aligned} L_{reg}(G) =& \mathbf{I}\lambda_{3D}L_{3D}(\Delta q|q^*_{3D}) + \\ & (1-\mathbf{I})\lambda_{2D}L_{2D}(\Delta q|q^*_{2D}) + \lambda_{bone}L_{bone} \end{aligned} \tag{2}$$

where $\mathbf{I}$ is the indicator function, which assesses whether the data are fully labeled data or weakly labeled data, and λ_{3D} and λ_{2D} are the loss weights.

3D Joint Loss. The joint loss L_{3D} enforces the predicted position of the 3D joints to be close to the ground truth position of the 3D joint. L_{3D} can be defined as follows:

$$L_{3D} = \|q^* - (q_{init} + \Delta q)\|^2 \tag{3}$$

where q^* is the ground truth 3D human pose, q_{init} is the predicted initial pose, and Δq is the predicted offset between the initial pose and the ground truth pose.

2D Joint Loss. We use 2D joint label supervision to learn the 3D joint location, and adopt a 2D joint loss $L_{2D}(\Delta q|q^*_{3D}, q^*_{2D})$, which can reduce the searching space of the 3D human pose. The 2D joint loss is defined as follows:

$$L_{2D} = \|q^*_{2D} - q_{2D}\|^2 \tag{4}$$

where q^*_{2D} is the ground truth 2D joint location, $q_{2D} = K(q_{init} + \Delta q)$ is the estimated 2D joint location and K is the camera intrinsic matrix.

Bone Length Ratio Loss. To penalize illogical bone length ratios, we add a bone length ratio regularization loss L_{bone} to generate the 3D human pose. The bone length loss L_{bone} indicates that the bone length ratio computed with the predicted joints is as close as possible to the average bone length in the training dataset

$$L_{bone} = \frac{1}{|E|}\sum_{e\in E}(\frac{l_e}{\bar{l}_e} - \bar{r})^2, \quad \bar{r} = \frac{1}{|E|}\sum_{e\in E}\frac{l_e}{\bar{l}_e} \tag{5}$$

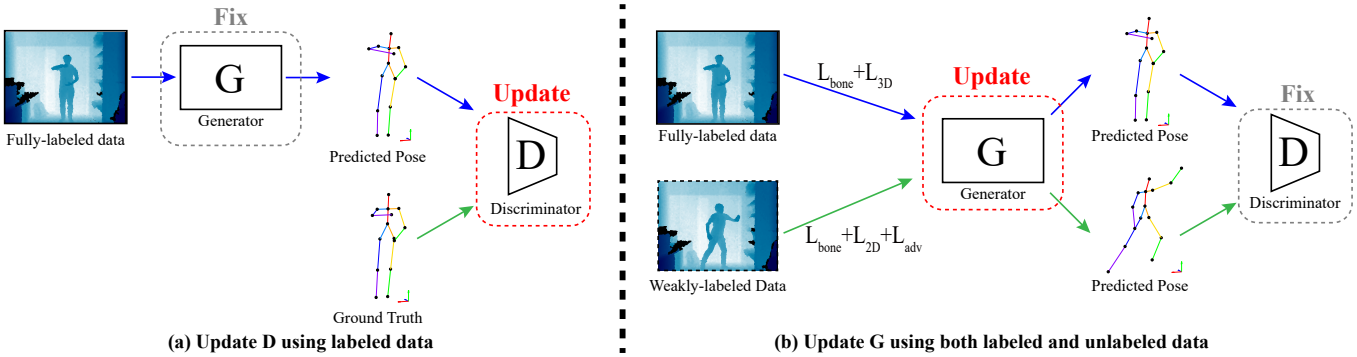


Fig. 5. Illustration of adversarial learning for the 3D human pose estimation. During the training stage, we adopt alternatively optimizations of the 3D pose regression and the discriminator. (a) Train the discriminator with the parameters of 3D pose generator fixed; (b) Train the 3D pose generator by simultaneously using fully labeled data and weakly labeled data.

where E is the set of all pairs of the bones in the used human skeleton model, $\bar{r}_i$ is the average bone length ratio between the predicted skeleton l_e and the average skeleton length $\bar{l}_e$ for the subjects in the training dataset. This term help to yield the human poses, the bone length ratio of which is close to that in the training dataset; thus, this term could enforce a more reasonable skeleton model.

3.3.3 Adversarial Learning

We adopt the 3D joint loss L_{3D} for the fully labeled data and the 2D joint loss L_{2D} for the weakly labeled data to learn the human pose estimation model along with the bone length ratio loss L_{bone}. The 2D joint loss forces the network to predict 3D joints that can be reprojected to the given 2D joint locations. However, it is generally ill-posed to reconstruct 3D human joints from 2D joints, i.e.many illogical 3D joint configurations can lead to the same 2D joints. To regularize the ill-posed issue, we adopt a discriminator network D to judge whether the reconstructed human joints are valid.

Fig. 5 illustrates the details of our human pose estimation network. The generator network is the 3D pose regression network, and the discriminator uses a fully connected neural network to distinguish the estimated human pose and the ground truth human pose. Denote q and q^* as the estimated and the corresponding ground truth human poses, respectively. The loss function of our network built upon a GAN can be formulated as:

$$L^G_{adv}(q,q^*) = \mathbf{E}_{q^*}[\log D(q^*)] + \mathbf{E}_q[\log(1-D(q))] \tag{6}$$

where $q = G(p_{norm})$ is the predicted human pose with input point clouds p_{norm}.

A generator G is trained to minimize the objective function (6) against an adversarial network D that tries to maximize it. We alternatively optimize G and D based on the minmax strategy

$$\min_G \max_D L_{reg}(G) + \lambda_{adv} L_{adv}(G,D) \tag{7}$$

where λ_{adv} is the weight of the adversarial loss.

Following the typical training procedure of the GAN, we alternatively update between discriminator D and generator G while fixing the parameters of the other network.

Updating Discriminator D. We train the discriminator D to classify between the ground truth 3D pose and the predicted 3D pose, which are labeled as 1 and 0, separately. Updating discriminator requires the ground truth 3D human pose; thus, we adopt the fully labeled data in this step. The optimization problem is equal to minimizing the binary cross-entropy loss $L_{BCE}(l^*,l) = -l\log(l^*) - (1-l)\log(1-l^*)$, where l^* is the output of the discriminator and l is the target label. To update the discriminator, we use the following loss function:

$$L^D_{adv}(q,q^*) = -\log D(q^*) - \log(1-D(q)) \tag{8}$$

Updating Generator G. To train the generator, the generated 3D human pose should be realistic enough to fool the discriminator. Therefore, the discriminator is trained to minimize $-\log(D(q))$.

To update the generator G, we utilize both the discriminator loss and the regression loss

$$\sum_{i\in\mathscr{S}_{full}} (L_{reg} + \lambda_{adv} L^G_{adv}) + \sum_{j\in\mathscr{S}_{weak}} \lambda_{adv} L^G_{adv} \tag{9}$$

where the adversarial loss L_{adv} is computed for both fully labeled data $\mathscr{S}_{full}$ and weakly labeled data $\mathscr{S}_{weak}$, which could benefit the training convergence on weakly labeled data.

3.4 Implementation Details

All our experiments are implemented within the TensorFlow framework on a workstation with two Intel Core i7 4790K processors, 32GB of RAM and an Nvidia Tesla K40 GPU.

The ITOP dataset [12] contains two kinds of data: clean, human-approved data with both 2D labels and 3D labels, and noisy human body-part labeled data with only 2D labels. We treat the clean data as the fully labeled dataset and the noisy data and the generated data as the weakly labeled dataset. In each mini-batch, we randomly sample 5 image pairs from both the fully labeled dataset and weakly labeled dataset. We conduct online data augmentation through scaling between $[1, 1.5]$ and rotating between $[-8°, 8°]$.

In our experiments, we find it more stable and effective to train the whole system in two stages. In stage 1, we initialize the 2D pose estimation module. For point clouds proposal stage, we set the width and height of each cropped patch to 20 pixels. In stage 2, we train our 3D pose regression module in a weakly supervised adversarial manner. Since the argmax operation is not differentiable, we did not train two stage network in an end-to-end manner. For each point-set abstraction level in hierarchical PointNet, we sample the point set from each patch and then reshape it into a holistic patch. This method could relieve the spatial context loss of 3D human joints in conventional random point sampling strategy [26].

We use Adam optimizer with a learning rate 0.0001 and a batch size of 10. The learning rate is set to decay 0.05% every 1000 iterations. The training process is stopped after 50 epochs. In our experiments, we set the weight λ_{3D}, λ_{bone}, λ_{2D} and λ_{adv} as 10, 1, 1e-3 and 1.

4 Experiments

In this section, we first introduce the used datasets and evaluation metrics, then conduct systematic evaluations of our methods: 1) self-comparisons and ablation study (shown in 4.2); 2) comparisons with the state-of-the-art (shown in 4.3), and show the qualitative results in Fig. 6.

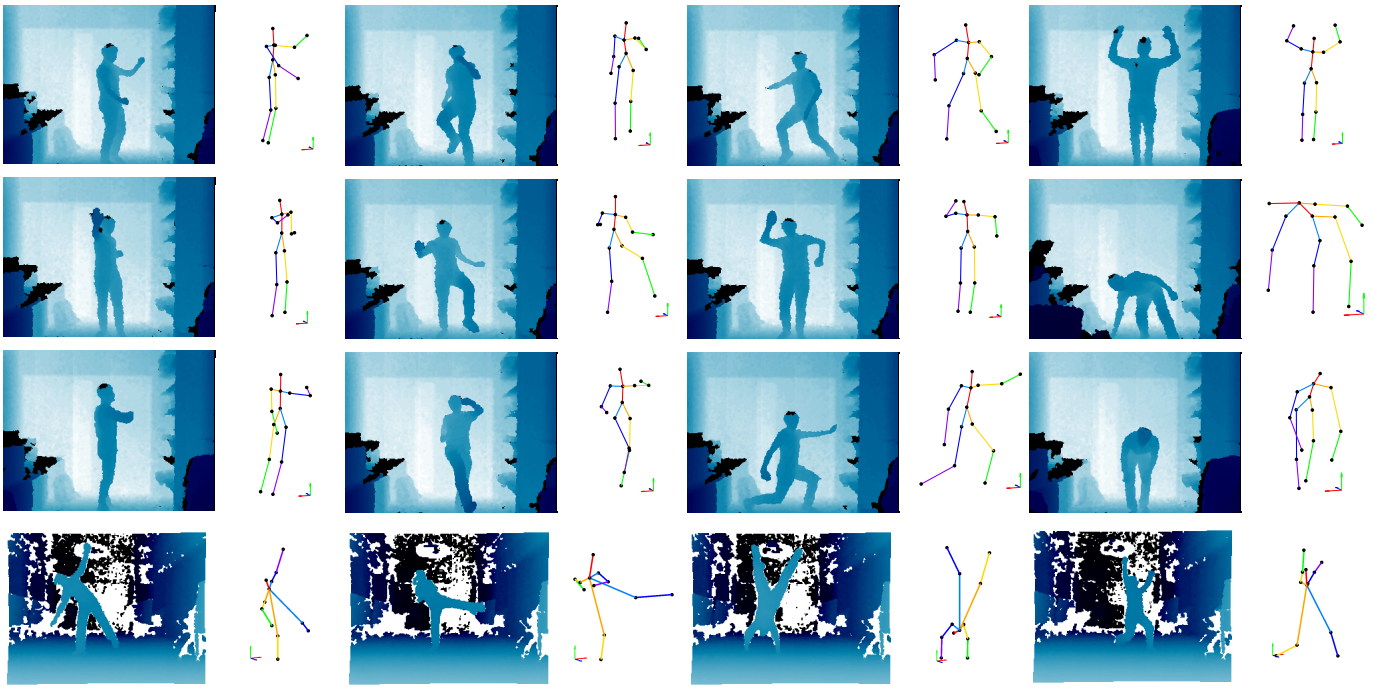


Fig. 6. Qualitative results from the ITOP dataset (the first three rows) and the EVAL dataset (the fourth row). The odd columns show the input depth map, and the even columns show our results. (The estimated 3D human pose may be not shown under the camera viewpoint)

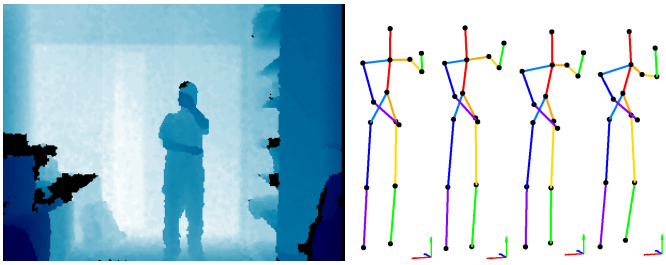


Fig. 7. Qualitative example of our self-comparison results. From left to right, we show (a) the input depth map, (b) the ground truth pose, (c) our result, (d) the result without bone length constraint, and (e) the result without weakly supervised learning.

4.1 Datasets and Metrics

In our experiment, we use the ITOP dataset [12] and Stanford EVAL dataset [8] to evaluate our method. The ITOP dataset contains more than 40K training samples and 10K testing samples from 20 subjects. Each subject has 15 actions. The dataset provides both side-view data and top-view data. The 3D pose ground truth contains $N = 15$ joints. We do not perform any data augmentation on this dataset during training. The EVAL dataset consists of 9K depth images from 3 subjects performing 8 different actions.

To evaluate the performance of our human pose estimation method, we use the following two standard metrics [19]. The first set of metrics are the percentage of correct keypoints (PCK) and mean average precision (mAP). The PCK value is the percentage of detected joints out of all human joints within a given distance error threshold. The mAP is the average PCK for all human body parts. It shows the overall robustness of the pose estimation methods. The other set of metric is the mean joint error, which is defined as the mean 3D distance error of each human joint. The mean joint error can indicate the accuracy of each joint.

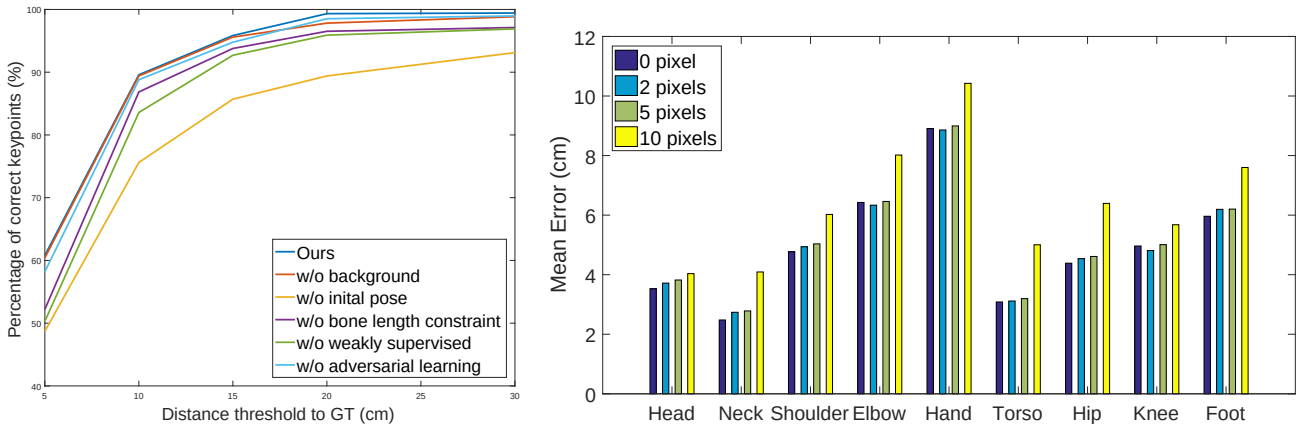
4.2 Self-comparison and Ablation Study

We conduct ablation studies to demonstrate the contribution of each component in our model and help us understand the design of our network better. We provide several examples under different self-comparison settings in Fig. 7 to visualize the effect of different components in our network. We use the ITOP dataset for the ablation study, and the detailed results are shown in Fig. 8 and Table 1. More evaluations are shown in the supplementary video.

Effect of 3D Initial Pose. To demonstrate the effect of the 3D initial pose for our 3D pose regression network, we compare the human pose performance with the 3D initial pose by directly predicting the 3D human pose by keeping the rest of our network fixed. As shown in Fig. 8(a), the mean average precision is 14 percentage points higher using the 3D initial pose at the 10 cm threshold, which indicates that the 3D initial pose is effective. We hypothesize that it is easier to learn the residue to the 3D initial pose than to directly learn the 3D human pose similar to the spirit of ResNet [13].

Effect of the Background. To demonstrate that our method is not affected by the background, we compare our method with the model trained without the background point clouds. For the model trained without the background, we skip the background during the point sampling process. As shown in Fig. 8, we observe that the PCK curves between the models with and without background points are similar. This demonstrates that the background has little effect on the robustness of our method. This result may be explained by the fact that the selected point clouds for 3D pose regression network are mainly around the initial joints and only contain few background points.

Effect of Bone Length Ratio Loss. To better understand the effect of our bone length ratio loss, we evaluate it in our pose regression network by removing the bone length ratio loss. As shown in Fig. 8, the accuracy with the bone length constraint is 2.8 percentage points higher than that without it at the 10 cm threshold. The mean joint error without the bone length ratio loss is approximately 0.53 cm higher than that with the bone length ratio loss. Fig. 7 shows a qualitative example of the impact of the 3D bone length constraint. We observe that the estimated left shoulder is of the wrong bone length compared with the result with bone length constraint and the ground truth. These results



(a) (b)

Fig. 8. The ablation study for different components in our method. (a) PCK by removing different terms in our method. (b) Mean 3D joint distance error with different levels of 2D pose errors in our method.

can be interpreted by the fact that bone length ratio loss can constrain the inter-joint constraint and minimize the occurrence of infeasible 3D human joints by enforcing the bone length ratio computed with the estimated joints.

	mAP	error (cm)
Our method	**89.59**	**5.51**
w/o background points	89.27	5.60
w/o initial pose	75.64	8.97
w/o bone length constraint	86.85	6.04
w/o weak supervision	84.58	6.68
w/o adversarial learning	87.20	5.95

Table 1. Detailed self-comparison results.

Effect of Weakly Supervised Learning. We also study the impact of the weakly supervised learning on our model. To evaluate the effectiveness of weakly supervised learning, we also conduct the experiments only using the fully labeled data to train the model and fixing the rest of the network. The results are shown on the left of Fig. 8. We observe that the our method with weakly supervised learning improves the mAP of the model without weakly supervised learning by approximately 6 percentage points at a threshold of 10 cm.

Effect of Adversarial Learning. We investigate the effect of adversarial learning by removing the adversarial loss and fixing the rest of the network. As shown on the left side of Fig. 8, the adversarial learning method helps to improve the mAP by approximately 0.8 percentage point higher at the 10 cm threshold.

Effect of 2D Pose Estimation Error. To demonstrate that our 3D regression network is robust to the error in the estimated 2D joint positions, we add random offsets to the estimated 2D joint positions to simulate the 2D joint estimation error, and compare the mean joint error under different offsets in the u-axis and v-axis of the image plane. We adopt four sets of offset parameters, i.e. 0 pixel, 2 pixels, 5 pixels and 10 pixels. As seen on the right side of Fig. 8, the performance of our network under a 5 pixel offset in each axis is still reasonably good. These results may be explained by the fact that the sampled point clouds could always contain the correct joint under this circumstance. However, when the 2D joint error is too large (see performance with 10 pixel offset), we cannot ensure that the sampled point clouds could contain the correct joint and its region of interest from the point clouds, and thus the performance may be poor.

4.3 Comparisons with State-of-the-art Methods

We evaluate the performance of our methods on the ITOP and EVAL datasets. During the comparison, we use state-of-the-art 3D human pose estimation methods, including the viewpoint-invariant feature-based method [12], the inference embedded method [32] and V2V-PoseNet [19]. The qualitative comparison results are shown in Fig. 9, and Fig. 10 shows the quantitative results on the ITOP dataset. Fig. 6 shows examples of the qualitative results, and more experiments are shown in the supplementary video.

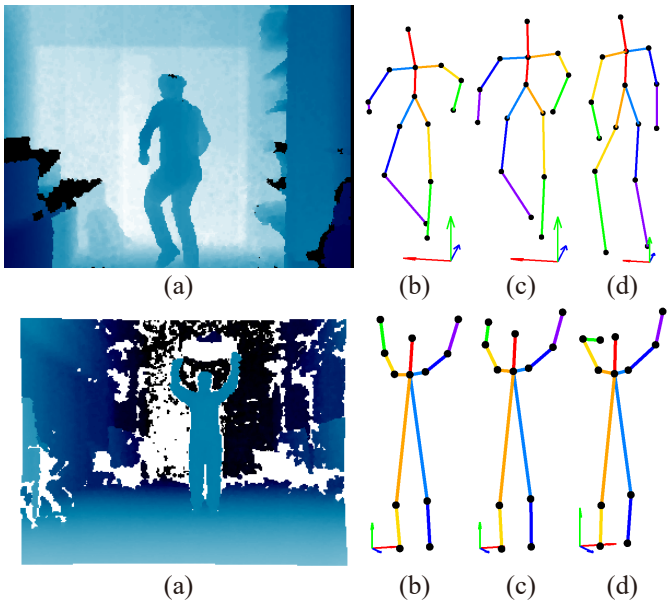
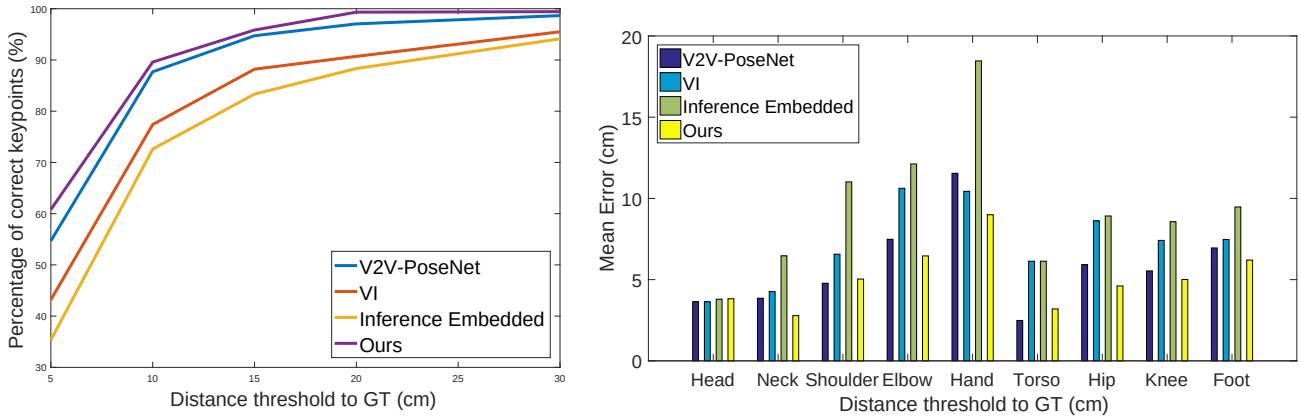


Fig. 9. Qualitative comparisons with the state-of-the-art methods.The first row shows results on the ITOP dataset, and the second row shows the results on the EVAL dataset. From left to right, we show the input depth map (a), the ground truth (b), the result of our method (c) and the result of the V2V-PoseNet method (d).

As shown in Fig. 10, our method can achieve state-of-the-art performance on the ITOP dataset compared with other 3D human pose estimation methods from point clouds. The detailed comparison results on the ITOP dataset and EVAL dataset are shown in Table 2.

For the ITOP dataset, the mean average precision with our method is 6.1 percentage points higher than that with V2V-PoseNet [19] if the threshold is set to 5 cm, and is 1.9 percentage points higher under the threshold of 10 cm. For the mean joint error, our method achieves better results compared with the state-of-the-art method V2V-PoseNet [19]. The average joint errors with our method are 2.7 cm,



(a) (b)

Fig. 10. Comparison of the proposed method with state-of-the-art methods. (a) mAP over different distance thresholds for different methods; (b) Mean 3D joint distance error for different methods. VI: viewpoint-invariant feature-based method [12].

	mAP (ITOP)				mAP (EVAL)	
Body part	[32]	[12]	[19]	Ours	[12]	Ours
Head	95.6	98.1	**98.25**	98.15	**93.9**	91.43
Neck	94.2	97.5	98.8	**99.47**	**94.7**	92.66
Shoulders	87.3	96.5	98.25	94.69	87	**88.21**
Elbows	72.5	73.3	78.73	**82.80**	45.5	**77.14**
Hands	53.8	68.7	67.21	**69.10**	39.6	**64.37**
Torso	85.4	85.6	98.29	**99.67**	-	-
Hips	70.5	72	90.25	**95.71**	-	-
Knee	64.2	69	**91.68**	91.00	83.4	**88.21**
Feet	58.8	60.8	85.87	**89.96**	**92.3**	**83.81**
Mean	72.62	77.4	87.69	**89.59**	74.1	**81.73**

Table 2. Comparison of joint mean average precision to state-of-the-art methods.

4.5 cm, and 1.3 cm lower than those with the viewpoint-invariant feature-based method [12], the inference embedded method [32] and V2V-PoseNet [19], respectively. In particular, we can achieve better results on the lower part of the body. The reason is that the accuracy of our method benefits from the robust 2D pose estimation process, and the 2D human pose estimation of lower body is overall better than that of the upper body. Furthermore, the 3D joint variance of the lower body in the ITOP dataset is lower than that of the upper body. Therefore, it could be easier for us to learn the 3D context of the point clouds and joints of the lower body and obtain reasonable results. Fig. 9 shows the qualitative results for the ITOP dataset. We can observe that our method can predict reasonable 3D human poses.

For the EVAL dataset, we compare our method with the state-of-the-art viewpoint-invariant feature-based method [12]. As shown in Table 2, the mean average precision (mAP) with our method is 7.6 percentage points higher than that with the method [12] if the threshold is set to 10 cm.

4.4 Runtime Analysis

We further investigate the efficiency of the proposed method. The training time of the ITOP dataset is 3 hours on a Tesla K40C graphic card. The testing time is 20 fps using the same GPU, and can be further increased in a multi-GPU environment. The detailed runtime comparison with state-of-the-art methods is shown in Table 3. During the comparison, we ran the models of state-of-the-art methods on the same workstation. We observe that our method is approximately 7+times faster than the other methods while achieving state-of-the-art accuracy.

Methods	FPS	mAP (ITOP)
Our method	**24.4**	**89.59**
[12]	0.6	77.4
[32]	7.4	72.62
[19]	3.5	87.69

Table 3. Comparison of runtime and accuracy performance with state-of-the-art methods

5 Conclusions

In this work, we adopt an efficient approach that exploits both the 2D and 3D representations of depth images or point cloud to achieve accurate 3D human pose estimation, and propose an effective weakly supervised adversarial learning method to learn 3D human pose estimation model using both the fully labeled 3D data and weakly labeled 2D data. Our weakly supervised method could relieve the lack of training data with 3D joint annotation in applications, and also inspire related researches such as human pose/shape recovery, in which full supervisions are not available or insufficient. Our experiments on the benchmark datasets show that our method can achieve the state-of-the-art performance.

Though inspiring results were obtained from this research, this work can be further improved to be an end-to-end framework. Specifically, our algorithm failed to fulfill the end-to-end prediction task due to the non-differentiable argmax operation in the 2D pose detection step. Although differentiable alternative operations such as soft argmax exist, they do not work well in our algorithm. It would be worthwhile to find a way to make our method end-to-end in a future work. Moreover, designing or utilizing other human pose representations would improve the potential of our method; for example, we can use human kinematics model with the joint angles and joint offsets and use them as the pose regression target, which could ensure that the reconstructed human joints are more plausible.

Our human pose estimation method can be used in several virtual reality applications that require the human body pose. Our method can help such applications to capture high-quality human body poses efficiently and enable more potential interactions, such as accurate trace tracking and fast-move tracking.

Acknowledgments

This work was supported by the Natural Science Foundation of Beijing Municipality No. L182052, National Key R&D Program of China under Grant 2016YFB1001201, National Natural Science Foundation of China No. 61772499 and No. 61473276, and in part by the Distinguished Young Researcher Program, Institute of Software Chinese Academy of Sciences.

REFERENCES

[1] A. Blum, A. Kalai, and H. Wasserman. Noise-tolerant learning, the parity problem, and the statistical query model. *Journal of the ACM (JACM)*, 50(4):506–519, 2003.

[2] Z. Cao, G. H. Martinez, T. Simon, S. Wei, and Y. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

[3] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[4] O. Chapelle, B. Scholkopf, and A. Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

[5] X. Chen and A. L. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *Advances in Neural Information Processing Systems*, pp. 1736–1744, 2014.

[6] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu. RMPE: Regional multi-person pose estimation. In *International Conference on Computer Vision*, 2017.

[7] P. F. Felzenszwalb, R. Girshick, D. Mcallester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[8] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun. Real-time human pose tracking from range data. In *European Conference on Computer Vision*, pp. 738–751. Springer, 2012.

[9] W. Gao, L. Wang, Z.-H. Zhou, et al. Risk minimization in the presence of label noise. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[10] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon. Efficient regression of general-activity human poses from depth images. In *International Conference on Computer Vision*, pp. 415–422. IEEE, 2011.

[11] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg. Matchnet: Unifying feature and metric learning for patch-based matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3279–3286, 2015.

[12] A. Haque, B. Peng, Z. Luo, A. Alahi, S. Yeung, and L. Fei-Fei. Towards viewpoint invariant 3d human pose estimation. In *European Conference on Computer Vision*, pp. 160–177. Springer, 2016.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[14] M. Kocabas, S. Karagoz, and E. Akbas. Self-supervised learning of 3d human pose using multi-view geometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1077–1086, 2019.

[15] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen. Pointcnn: Convolution on x-transformed points. In *Advances in Neural Information Processing Systems*, pp. 820–830, 2018.

[16] M. J. Marin-Jimenez, F. J. Romeroramirez, R. Munozsalinas, and R. Medinacarnicer. 3d human pose estimation from depth maps using a deep combination of poses. *Journal of Visual Communication and Image Representation*, 55:627–639, 2018.

[17] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2640–2649, 2017.

[18] D. Maturana and S. Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 922–928. IEEE, 2015.

[19] G. Moon, J. Yong Chang, and K. Mu Lee. 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*, pp. 5079–5088, 2018.

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

[21] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7025–7034, 2017.

[22] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Harvesting multiple views for marker-less 3d human pose annotations. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 6988–6997, 2017.

[23] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 918–927, 2018.

[24] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. 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.

[25] C. R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, and L. J. Guibas. Volumetric and multi-view cnns for object classification on 3d data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5648–5656, 2016.

[26] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, pp. 5099–5108, 2017.

[27] 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*, pp. 8437–8446, 2018.

[28] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1297–1304. Ieee, 2011.

[29] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[30] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei. Integral human pose regression. pp. 536–553, 2018.

[31] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1653–1660, 2014.

[32] K. Wang, S. Zhai, H. Cheng, X. Liang, and L. Lin. Human pose estimation from depth images via inference embedded multi-task learning. In *Proceedings of the 24th ACM international conference on Multimedia*, pp. 1227–1236. ACM, 2016.

[33] P. Wang, Y. Liu, Y. Guo, C. Sun, and X. Tong. O-cnn: octree-based convolutional neural networks for 3d shape analysis. *ACM Transactions on Graphics*, 36(4):72, 2017.

[34] 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*, 2016.

[35] S. Xia, Z. Zhang, and L. Su. Cascaded 3d full-body pose regression from single depth image at 100 fps. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 431–438. IEEE, 2018.

[36] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu. Pose Flow: Efficient online pose tracking. In *British Machine Vision Conference*, 2018.

[37] C. Zhang, J. C. Platt, and P. A. Viola. Multiple instance boosting for object detection. In *Advances in Neural Information Processing Systems*, pp. 1417–1424, 2006.

[38] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 398–407, 2017.

[39] Y. Zhou and O. Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4490–4499, 2018.

[40] Z. Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2018.