BoneNet: Real-time 3D Human Pose Estimation By Generating Multiple Hypotheses with Bone-map Representation

Guodong Wei[1] , Shihao Wu[2] , Keke Tang[3] and Guiqing Li[1]

[1]South China University of Technology, g.d.wei.china@gmail.com,ligq@scut.edu.cn
[2]Swiss Federal Institute of Technology Zurich, shihao.wu312@gmail.com
[3]Guangzhou University, tangbohutbh@gmail.com

Corresponding author: Guiqing Li, ligq@scut.edu.cn

Abstract. A robust and fast 3D human pose estimation technique is essential for tasks like geometric modelling of human, human-computer interaction, and augmented reality. In this paper, a fully convolutional neural network is developed for extracting 2D and 3D human poses from a single image. Based on the observation that the existing 3D pose representation, called joint location map, fails to handle cases where occlusions occur, the new 3D human pose representation named bone map is thus proposed in this work. The key idea includes two aspects: first, the network generate a set of independent pose hypotheses, each of which is associated with a heat map indicating the location of a 2D keypoint, so a complete 3D pose can be read out from any 2D joint that is not occluded; second, hyper-bone maps are derived from bone maps, to provide stronger supervision to the 3D pose structure and thus enhance the network ability of predicting the locations of missing joints. By jointly optimizing a well-organized set of 2D heat maps and 3D bone maps, a portable network is trained and applied in real time to address challenging cases where occlusion occurs. Qualitative and quantitative experiments show that our model achieves comparable performance to the state-of-the-art 3D pose estimation methods.

Keywords: monocular 3D human pose estimation, real-time, 3D pose representation, fully convolutional network
DOI: https://doi.org/10.14733/cadaps.2021.1448-1465

1 INTRODUCTION

Human pose estimation has been extensively studied due to its critical role in geometric modelling of human, human-computer interaction, and augmented reality, etc. Taking the human modelling as an example, 2D or 3D skeleton poses often serve as constraints that drive the deformation of a morphable model of human body [6, 15, 21, 34, 37]. Traditionally, human poses can be obtained by commercial motion capture systems

that require controlled environments and expensive equipment, such as landmark trackers and stereo depth cameras. Recently, more and more researchers focus on extracting 2D and 3D human poses from 2D images so as to make the human modelling easier and flexible.

Unfortunately, reconstructing the 3D pose from a single RGB image is an ill-posed problem in nature. The challenge comes from the data ambiguity, such as self-occlusion, inter-object occlusion, and complex illumination. Early work addresses this problem in a simplified setting [1, 38, 56], or leverages additional information [5, 7]. With the emergence of deep neural networks, plenty of deep learning based methods have achieved significant improvement in 2D human pose estimation [8, 32, 47, 50, 53]. The success of data-driven approaches in 2D pose estimation encourages the development of deep neural networks for image-based 3D pose prediction.

The heat map representation, which stores the probability of a joint's presence at each and every pixel in the image, has dominated the neural 2D pose estimation, thanks to the power of fully convolutional networks. In deep 3D pose prediction, however, a two-step strategy is often used. Namely, it first estimates the 2D pose and then regresses the 3D posture based on the 2D joint prediction [6, 9, 25, 30, 33, 46, 52, 55, 58]. The output of the first step is a sparse set of 2D joint predictions and it maybe relatively unreliable, this makes the second step probably missing some useful image cues and results in erroneous 3D pose reconstruction. On the other hand, researchers also attempt to directly regress for the 3D pose by using deep networks. These networks typically consist of convolutional layers for feature extraction and several linear layers for the pose regression. However, the performance of direct regression methods is generally worse than those two-step methods. This is probably because of that regression learning is more difficult than the densely supervised heat map learning [43]. And the use of linear layers also introduce a large amount of parameters which may cause over-fitting and additional computational overhead.

To this end, we propose a novel 3D human pose representation called bone map and devise a fully convolutional network i.e., BoneNet, for fast and accurate 3D pose estimation. Intuitively, the human posture can be represented as a set of bones, i.e., line segments. Each bone can then be simplified to a vector (x, y, z) given the root position and pre-defined bone connectivity. Then this vector is encoded into three bone maps which have the same dimensionality as heat maps. The key technique includes three aspects: first, each bone map stores the possible values of the coordinate rather than the probability of the joint location; second, each heat map is associated with an independent 3D pose, i.e., a set of bone maps; third, the bone connectivity is sustained by using *hyper-bone maps*.

A loss function, which is used to supervise the network in training stage, is also carefully designed to ensure both types of maps to be compatible with each other. The output of our pose estimation network is a set of 3D skeletons and a 2D skeleton. Based on these predictions, two approaches are proposed to extract the final 3D pose: selection-based and regression-based. Although our method also consists of two steps, but the output of our first step is a set of 2D and 3D pose hypotheses, which should provide more contextual information and is clearly different with the aforementioned two-step methods that only generate one 2D pose at the first stage.

Our method is tested on two commonly used large-scale datasets. Qualitative and quantitative experiments show that our method can yield 3D pose predictions comparable to existing state-of-the-art methods in real time.

In summary, the main contributions of this paper include:

- The bone map is presented as a new representation of 3D human poses, which is compatible with the heat map; hyper-bone maps are also derived to provide effective supervision to the 3D pose predictions.
- Our 3D pose estimation network generates multiple 3D pose hypotheses from a single image, offering more space for the network to deal with the pose ambiguity problem.
- Two simple yet efficient techniques were explored for determining the final 3D pose and achieve relatively high estimation accuracy in real time.

2 RELATED WORK

The aim of this work is to develop a real-time 3D human pose estimation method for ordinary consumers. Commercial solutions usually require expensive equipment, a controllable environment, and professional skills. Interested readers are referred to [10, 41] for a survey of these approaches. Here, the rest of this section only focus on recent works based on regression and detection.

2.1 Detection-based Methods

Detection-based methods have been the main driver behind the recent development of 2D pose estimation [8, 32, 47, 50, 53]. These methods are based on heat maps in which the pixel having the largest confidence is a specific 2D joint located. Nevertheless, extending this kind of methods to 3D pose estimation is not straightforward, due to the inherent difference of embedding space of 2D and 3D poses. Pavlakos *et al.* [36] address this problem by extending 2D heat maps to 3D via a volumetric representation, and a coarse-to-fine prediction scheme is proposed to reduce memory consumption. Fabbri *et al.* [13] devise a compression method to drastically reduce the size of volumetric heat maps and apply it to multi-person 3D pose estimation. Even so, it is still a huge model with heavy computation due to the use of volumetric heat maps. Besides, the discrete characteristics of 2D and 3D heat maps make the model suffer from quantization errors. The quantization errors occurs when the continuous coordinates of human joint locations are quantized and represented in heat maps with only discrete pixel locations. The joint localization precision is thus limited by the quantization factor or the resolution of the heat map.

2.2 Regression-based Methods

Regression-based methods typically produce continuous output in the form of a vector, avoiding the quantization problem. The representation of 3D skeletal poses is an important factor in deep learning based methods and has attracted much attention of researchers. Some existing representations include root-centered joint positions [19], kinematic skeleton models [57] and relative joint positions [23]. Regression-based methods usually adopt the one-step strategy and are typically inferior to the two-step approaches that lift the predicted 2D joints to 3D.

Traditionally, the two-step approaches cast the 3D pose reconstruction problem into a complex nonlinear optimization with constraints like bone length [31], joint limits [2], inter-penetration [6], sparsity [49, 58], and temporal dependency [39]. The computation overhead is rather expensive. Recent work utilizes deep neural networks to reconstruct 3D pose [6, 9, 25, 30, 33, 35, 46, 52, 55, 58], which are usually faster and more effective. However, the reconstruction step heavily relies on the 2D prediction which inevitably introduces quantization errors and loses plenty of contextual information in the original image. Lin *et al.* [24] capture environment information by an intermediate module to encode 2D features into a vector and combine it to regress the 3D pose. Tekin *et al.* [44] suggest that 2D and 3D features can be fused optimally by using a trainable fusion scheme. Noticing that the vector-based bone representation [23] is proven to be useful in 3D pose prediction, Sun *et al.* [42] unify both 2D and 3D pose vectors using a compositional loss function. Different with them, our method extend the vector-based bone representation to a matrix-based bone map representation and apply it in a fully convolutional network that generates heat maps at the same time.

2.3 Combination of Detection and Regression based Methods

Combining the strengths of both detection and regression-based methods is an attractive direction. Sun *et al.* [43] propose an integral regression method that unifies the heat-map representation and joint regression. Mehta *et al.* [29] extend the 2D heat map to a set of 3D location maps. The authors associate each heat map with X, Y, and Z location maps, representing the corresponding 3D joint. This one-to-one association

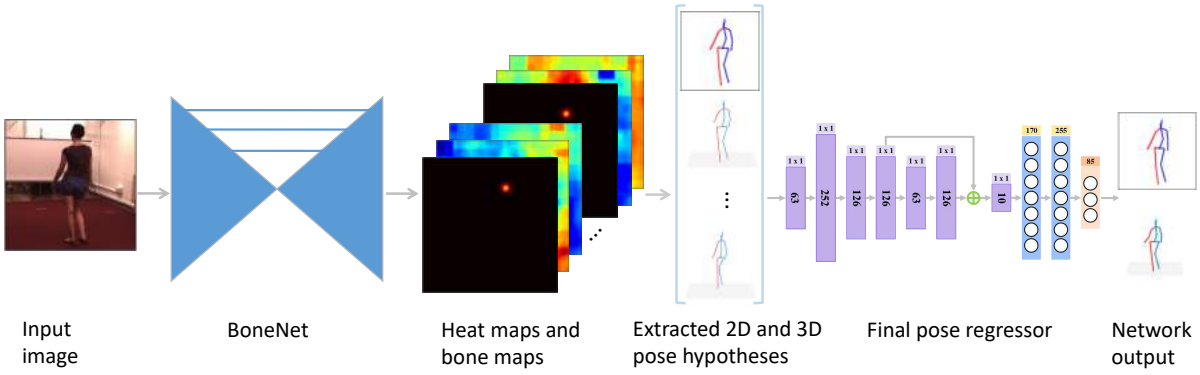

Figure 1: The overall pipeline of our method. The input is firstly fed into the BoneNet to generate the heat maps and bone maps. Then 2D joint positions are obtained by finding the locations of highest values in each of the heat maps, and the 3D joint positions are read out from bone maps at the same location of each corresponding 2D joint. These 2D and 3D joint positions are then fed into a pose regressor or a selection-based algorithm to get the final predictions. In this figure, only the pose regressor is shown.

between 2D and 3D joint prediction easily suffers from false predictions due to the occlusion. Later, Mehta *et al*. [27, 28] improve the 3D location map representation for the multi-person setting and address the problem mentioned above to some extent. In this paper, we propose a one-to-many approach for single person 3D pose estimation in which each heat map associates to a full skeleton represented by multiple bone maps to leverage the advantages of both local joint detection and global shape regression.

2.4 Methods Addressing Pose Ambiguities

Given a single image, there exist many reasonable 3D poses agreeing with the same 2D projection. This makes the 3D pose estimation to be an ill-posed optimization problem. This ambiguity can significantly affect the performance of the predictor [14]. Fan *et al*. [14] use random vectors to perturb the prediction, which enables their network to generate multiple plausible shapes from one input image. On the other hand, Pavlakos *et al*. [35] adopt weak supervision to 3D pose prediction by using ordinal depth to resolve the depth ambiguity of a single image. Zhou *et al*. [55] summarize that incorporating weakly supervised constraints is a powerful strategy for performance boosting. Li *et al*. [22] propose a novel approach to generate multiple feasible hypotheses of the 3D pose from 2D joints based on a multimodal mixture density network. Our method addresses the pose ambiguity problem in a different way, which produces pixel-wise 3D pose predictions for a single image.

3 METHODS

Given an RGB image I of a person, the goal is to generate 2D and 3D poses that best represent the human motion at that moment. However, this task is particularly challenging due to intra-object and inter-object occlusions. In this section, a learning technique that jointly estimates 2D joints and 3D bone locations will be described. The core of this technique is the bone-map representation of a 3D human skeleton and how it associates to 2D heat maps. Intuitively, each 3D skeleton is represented by a set of bone vectors with each vector encoded in three bone maps. Then each 2D heat map (encoding a 2D joint location) is associated to a group of bone maps that contain the full information of a human skeleton. In other words, our network predicts a single 2D pose (heat maps) and multiple 3D poses (groups of bone maps) from a single input image.

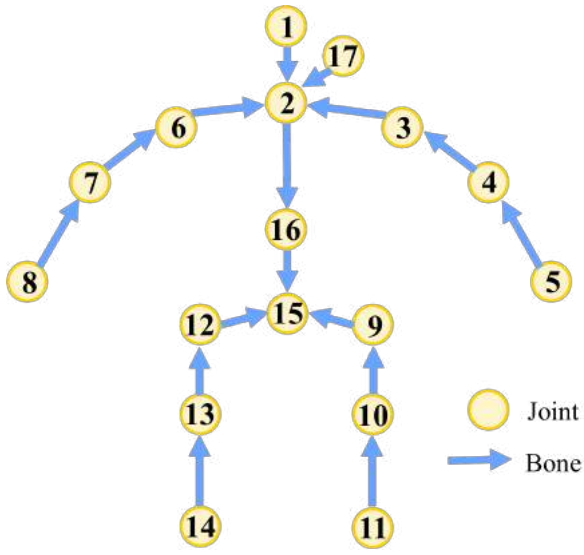

Figure 2: Structure of the human skeleton, including 17 joints J and 16 bones B. The pelvis (No. 15) is used as our root joint.

This can be viewed as a relaxation of the optimization problem, giving the network more chance to avoid the local minimum.

Our method encourages communication between global and local information by exploiting the power of convolutional operations on our compatible data representation. An overall pipeline of our approach is shown in Figure 1.

3.1 Pose Representation

Specifically, $P_{2D} \in \mathbb{R}^{K\times 2}$ and $P_{3D} \in \mathbb{R}^{K\times 3}$ will be estimated, from which 2D and 3D locations of K joints can be extracted respectively. Typically, the coordinate unit is pixel for 2D locations and millimeter (mm) for 3D locations.

Following the common practice, the heat-map representation is adopted for 2D pose estimation. Concretely, our network predicts a set of heat maps denoted by $H = \{H^i, i = 1, ..., K\}$ and each pixel in H^i stores the confidence value indicating whether joint i locates on it.

Inspired by the location maps [29], the bone map to represent a 3D pose is introduced. As shown in Figure 2, each bone of the human skeleton is represented as a *vector* pointing from a child joint to its parent joint, except for the root whose parent is defined as itself. The k_{th} bone vector is denoted by $b_k = (x_k, y_k, z_k)$. For each bone vector, our network predicts three bone map matrices $B_k = \{X_k, Y_k, Z_k\}$ which have the same dimensionality as heat map H_i. Instead of storing probability, each element in a bone map (X_k, Y_k, Z_k) corresponds to a possible value of the coordinate (x_k, y_k, z_k). Such a redundant mapping (64×64 to 1) not only offers sufficient network capacity for encoding the 3D coordinates without discretizing the 3D space, but also support establishing connection with 2D heat maps.

One of our key ideas is to make every 2D joint prediction contribute to predicting an entire 3D pose. It greatly improves the global-local consistency. More concretely, a 2D heat map H_i is associated with a group of 3D bone maps $B^i = \{B^i_1, B^i_2, ..., B^i_K\}$, where each group of bone maps stores the complete information of a 3D pose independently. The values of a different group of bone maps are tightly related to their corresponding

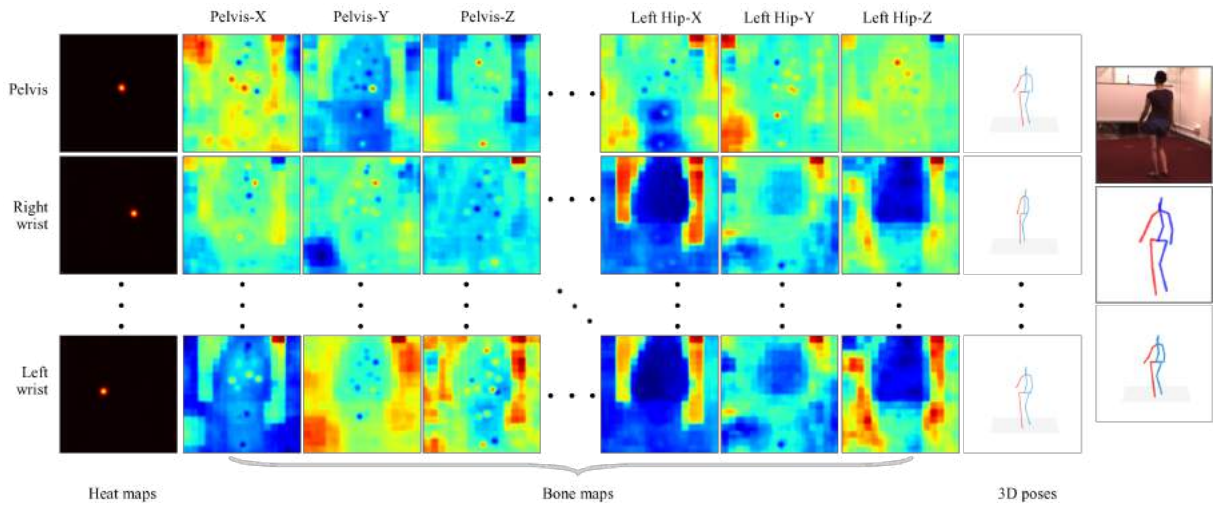

Figure 3: The visualization of heat maps and bone maps generated by our fully convolutional network. As there are $K = 17$ joints in the human model used in this paper, 17 heat maps and 17x17x3 bone maps are generated by the convolutional network. These heat maps and bone maps are placed into a matrix with 17 rows and $52 = 1 + 17 \times 3$ columns for visualization. A different row then contains one heat map and 17×3 bone maps corresponding to a specific kind of joints. In this figure, for each of the columns from left to right are: heat maps of the 2D joints, different groups ($K \times K \times 3$) of bone maps, the 3D poses extracted from each group of bone maps, and the last column from top to down shows the input image, the final 2D and 3D pose predictions based on the heat maps and bone maps.

heat map. The relation between heat maps and bone maps is illustrated in Figure 3.

By linking the heat map and bone maps, the 3D pose prediction is encouraged to consider the 2D information such that the 3D prediction is more consistent with the image. On the other hand, the 2D prediction can leverage the pre-defined structural knowledge, and the global information learned from 3D pose prediction, with the 3D bone maps stacked over each 2D heat map. This is also related to the attention mechanism that is widely used in neural image processing methods [11, 48, 51]. Here, the heat maps is used to direct the attention of the network to different local areas while predicting a global structure.

Besides boosting 2D-3D communication, our bone-map representation enables the use of convolutional layers rather than fully-connected layers for a 3D pose estimation. We observe that a bone of a human usually occupies a small region of an image, which arguably can be handled more efficiently by convolutional operations. This also suggests that our 3D prediction task is compatible with 2D one, i.e., both aiming to detect local subjects.

Although convolutional neural networks are strong in detecting local features, a 3D predictor requires global structural information, especially when occlusion occurs. Our method addresses this problem by extracting a full 3D pose from each of the 2D joints. In case a 2D joint is occluded and thus wrongly predicted, the prediction of the linked groups of bone maps has little impact on those unlinked groups. Soft constraints between bone maps and heat maps can bridge the local joint positions and the global structure. And it also allows performing structure constraint as described in Section 3.2. In the end, K 3D poses generated by our network may differ from each other, and each of them is an approximate prediction of the specific 2D joints.

3.2 Loss Functions

In this section, a hyper-bone map representation is firstly introduced to leverage the connectivity information between bones, and the loss functions defined upon hyper-bones is described.

Computing loss on bone maps is sub-optimal as the prediction of each bone map is independent, although a group of bone maps is assigned to each joint. In addition, the error of the individual predictions would propagate along with the skeleton and accumulate at the far end [42]. the prior knowledge of the human skeleton structure is exploited by constraining hyper-bone maps defined by the connectivity of all pairs of joints in the loss function. Concretely, a hyper-bone map $B_{u,v}$ is introduced for an arbitrary pair of joints (u, v) in the skeleton with each element representing a vector pointing from u to v. $B_{u,v}$ has the same dimension as B_k. Given the network prediction $\{H^i, B_k^i, ..., B_k^i, i = 1, ..., K\}$, i.e., the heat map and its corresponding bone maps. We transform the 3D bone maps into root-centered joint location maps $\{J_1^i, ..., J_K^i\}$ using following formula,

$$J_k^i = T(B_k^i) = \sum_{m \in Path(k)} B_m^i, \tag{1}$$

where $Path(k)$ is a function returning the set of bones belonging to the path from the joint k to the root. The hyper-bone map $B_{u,v}$ can then be calculated as

$$B_{u,v}^i = J_u^i - J_v^i. \tag{2}$$

The above analysis leads to the following loss function,

$$Loss_{3D} = \sum_{i<=K} \sum_{(u,v)\in P} \left\| \bar{H}^i \odot (B_{u,v}^i - \bar{B}_{u,v}^i) \right\|_1, \tag{3}$$

where $P = \{(u, v)|1 <= u < v <= K\}$ denotes all possible joint pairs, $\bar{H}^i$ and $\bar{B}_{u,v}$ are derived from the ground truth, and $\odot$ is the Hadamard product. Here, $L1 - norm$ is adopted for robustness against outliers that caused by occluded body parts.

Note that a great advantage of jointly optimizing heat maps and bone maps, e.g., the element-wise production in our loss function, is to unify the value distribution of a group of bone maps associated with a certain heat map. This enables us to treat a bone map as a regular scalar and makes the $+$ and $-$ operation of two in-group bone maps meaningful. In other words, our hype bone map, subtraction of two bone maps, can well represent a bone structure of a human skeleton. Since the bone vectors are physically linked with each other, the joint optimization can channel the attention of our network to both global and local contextual information.

The loss of 2D pose prediction is relatively straightforward, which is the mean square error between the predicted heat maps and the ground truth and is calculated as,

$$Loss_{2D} = \sum_{i<=K} \left\| H^i - \bar{H}^i \right\|_2. \tag{4}$$

Finally, the total loss is a compromise of the above two losses

$$Loss = Loss_{2D} + \alpha Loss_{3D}, \tag{5}$$

where α is a scalar that controls the weights balancing the losses of 2D and 3D predictions.

The poses in training data are normalized using standard normalization. The mean and standard deviation of each bone is firstly computed over the entire training set and then each bone position is normalized as,

$$\widehat{B} = \frac{B_i - mean(B_i)}{\lambda \cdot var(B_i)}. \tag{6}$$

It is observed that the scaling factor λ is rather important and $\lambda = 6.0$ is adopted in our experiments.

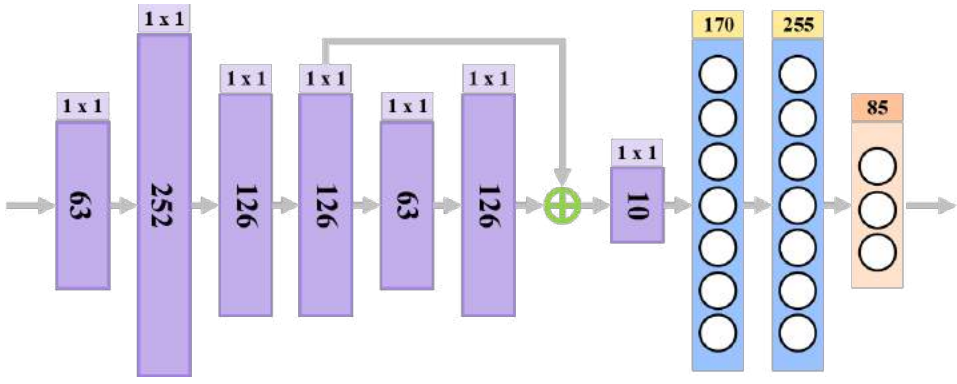

Figure 4: The structure of our pose regression network. The kernel size of a convolutional operation or the output feature size of a linear operation is shown at the top of each rectangle representing a module, while the number of output channel of a convolutional operation is shown inside.

4 Synthesize the Final Prediction

The network being discussed so far outputs K candidates of the 3D pose. A final 3D pose prediction needs to be generated based on the candidates. In this session, two options for determining the final pose will be discussed.

4.1 Selection-based Method

The selection-based method chooses one prediction from K candidates that best agrees with the 2D pose prediction. The selection is based on a metric related to the projection errors with respect to 2D prediction and the space-time smoothness prior. More specifically, the final prediction $\hat{P}$ should satisfy

$$\hat{P}_{3D} = \arg\min \left\| \Pi(P^i_{3D}) - P^i_{2D} \right\|_1 + \omega \|\widetilde{P}_{3D}\|_2, \tag{7}$$

for $i = 1, ..., K$, where Π is the projection matrix and $\widetilde{P}_{3D}$ is the average acceleration of joint movement. We use a scale weight ω to balance the two terms.

This selection based method has the following advantages. First, it leads to a small solution space and the optimization can be highly parallelized. Actually, the complete pipeline can run in **real time (20HZ)**. Second, the selection process does not require modifying the predicted poses and maintains the structure-aware output of the network.

The accuracy of our method is lower than that of the min-MPJPE (defined in Section 5.3) by a margin of about $5mm$. This may be caused by the inaccuracy of the 2D prediction, leading to a sub-optimal selection of 3D pose. Nevertheless, the performance of this simple selection method is close to the state-of-the-art methods.

4.2 Regression-based Method

An alternative to determining the final pose estimation is to employ a simple neural network for regression. It seems that this approach tends to yield a more accurate result and is more efficient than those existing selection-based methods.

The input of the regression network is formed by gathering the predictions of our Convolutional network on the Human3.6M dataset and then extracting 3D joint coordinates from them. The ground-truth of the

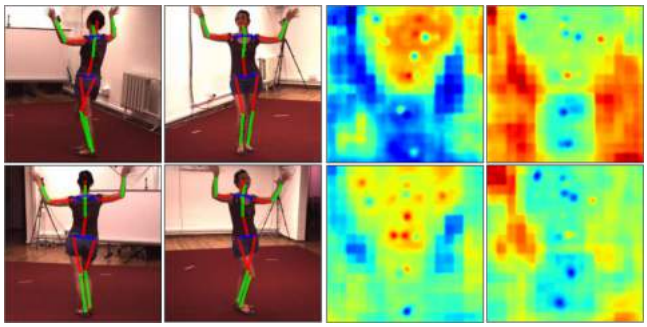

Figure 5: A visualization of four corresponding bone maps associate to the same joint (pelvis-X) from four different views. Rich information are stored in these bone maps. Interestingly, the changes of the bone map pattern reflect the updates of 3D orientation, even though the 2D poses can be quite similar (top-bottom pairs), indicating the bone map's ability to encode global structural information.

input is also collected correspondently from Human3.6M. During training, for arbitrary given frame $\mathcal{F}$, the estimated the 3D pose from the K predictions together with the 2D predictions is stacked into a matrix of $(3K+2) \times K \times 1$, where $(3K+2)$ is the number of channels and the $K \times 1$ features store values of x, y or z coordinates of the 2D or 3D joints.

As for the loss function, $L1 - norm$ is employed to compute the error between the predicted and the ground truth joint positions,

$$Loss_{reg} = \left\| P_{3D} - \bar{P}_{3D} \right\|_1 + \left\| P_{2D} - \bar{P}_{2D} \right\|_1 . \tag{8}$$

A deep network-based pose regressor is designed to takes the contextual matrix as input and predict our final 2D and 3D poses of frame $\mathcal{F}$. The network consists of a sequence of 1x1 convolution layers, along with a few fully connected layers. Some skip connections are also used in our network. The network architecture is shown in Figure 4. The Adam optimizer is used in optimization and learning rate is set to $6e-5$. The mini-batch size is 128. The training process is high-speed, and the model can converge within one hour due to a compact data representation and lightweight network design. Most importantly, the inference of our full model runs in **real time (30HZ)**.

On the Human3.6M test set, the reconstruction accuracy of our regression-based method is about $59.6mm$, which is slightly better than the min-MPJPE. For the consideration of speed, only the compact prediction of joints positions extracted from our BoneNet is utilized as the input of our regression-based pose synthesis method. It would be interesting to investigate the use of heat maps and bone maps as input, which will provide more information (Figure 5) and may lead to higher reconstruction accuracy.

5 EXPERIMENTAL RESULTS

In this section, a variety of experimental evaluations on the proposed approach are conducted. First, the benchmark datasets used for quantitative and qualitative evaluation are introduced . Then, some essential details for implementing our approach are provided. Finally, quantitative and qualitative results are presented on the selected datasets.

5.1 Implementation Details

As shown in Figure 6, our convolutional pose prediction network makes use of an encoder-decoder structure. It first downsamples the image to a small bottleneck for a large reception field and then upsamples the image to

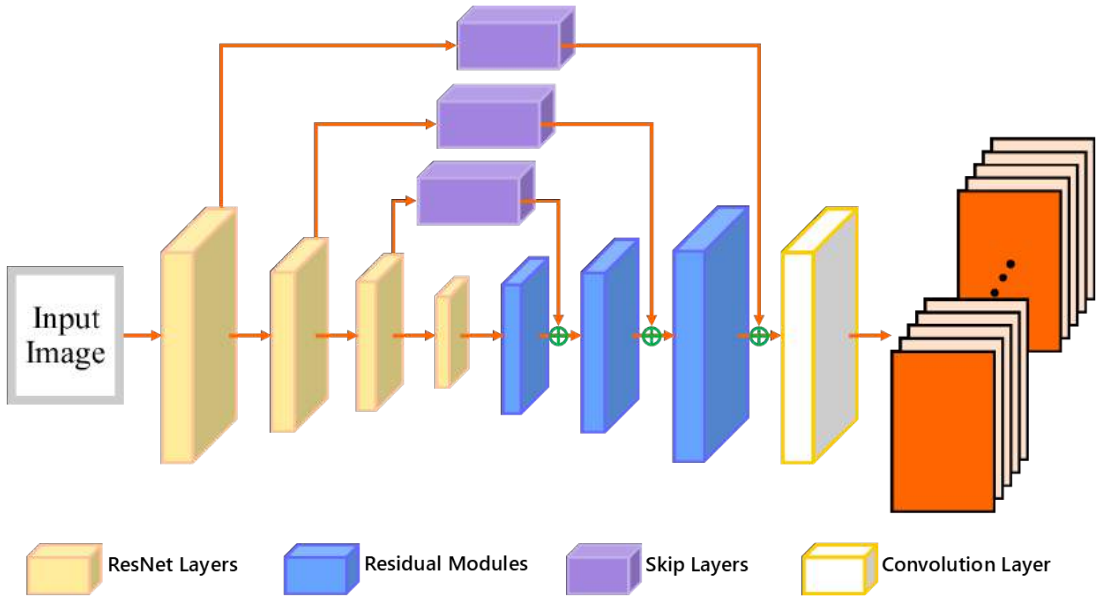

Figure 6: The architecture of our network (the BoneNet), which is an extension of ResNet50 [17]. Following the convention in the Hourglass network [32], a few skip layers between some lower-level and high-level feature maps are added. The output of our network is a set of heat maps for 2D joint positions and a set of bone maps for 3D bone positions. Each 2D heat map is associated with a group of 3D bone maps.

obtain fine-grained results. The encoder is built on ResNet50 [17]. Specifically, the trained ResNet50 model with its last two layers removed is used as the backbone. The resulting network then contains four parts: Layer1, Layers2, Layer3, and Layer4. Our decoder contains three up-sampling layers and a 1×1 convolutional layer. Inspired by the Stacked Hourglass Network [32], a skip layer between each pair of the encoder part and its corresponding decoder part of ResNet50 is added to facilitate residual learning. Note that, instead of directly computing fixed 3D joint coordinates, our bone-map representation can serve as a soft intermediate layer for residual learning, which has proven to be useful for many vision learning tasks. For regularization, two dropout layers are also adopted after each of the last two parts of ResNet50. The input image resolution of our network is 256×256, and the output is a set of 64×64 maps including 17 heat maps and $17 \times 17 \times 3$ bone maps. We use RMSprop as our optimizer with a batch size of 20, and a learning rate of 5e-5.

5.2 Datasets

Experiments are conducted on three popular 2D or 3D human pose estimation benchmarks: the Human3.6M [20], the MPI-INF-3DHP [26] and MPII [4] datasets.

Human3.6M [20] is, by far, one of the biggest datasets for 3D human pose estimation. It has 3.6 million annotated images, including 11 actors performing 15 daily activities from different views in an indoor environment. Following a standard protocol, subjects S1, S5, S6, S7 and S8 are used for training and subjects S9 and S11 for testing (Protocol #1). The original videos have been downsampled from 50Hz to 10Hz to reduce redundancy.

MPI-INF-3DHP [26] is a recent 3D dataset captured in both indoor and outdoor environments. The 8 training subjects from 8 cameras are adopted following [29]. To reduce redundancy, the training data is down-sampled such that at least one joint moves by more than 200 mm between consecutive frames. During

	min-MPJPE
Single Output	82.0
Multiple Output	68.2
Multiple Output + Hyper-bone	63.3
Multiple output + Hyper-bone + MPII	60.0

Table 1: Ablative study on the Human3.6M dataset demonstrates the effectiveness of our bone map based formulation and the hyper-bone map supervision.

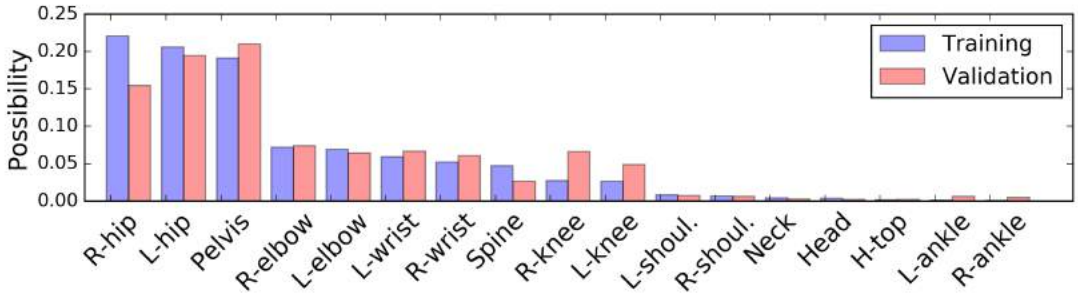

Figure 7: Study on the possibilities of the 3D pose prediction with minimal error located at each joints. Since one 3D pose hypothesis is stored at each of the 2D joint locations and only one 3D prediction is generated by our method, this study is to provide more insights about how these different 3D pose hypotheses contribute to the final prediction. The joints are sorted in descending order from left to right according to their possibility of relating to the final 3D pose in training.

training, the sampled images are augmented by random color jittering, different combinations of occluders, backgrounds, and clothing. To demonstrate the generalization of our model, the test set of the original dataset, which contains 2929 frames of six subjects performing seven actions, is also used in evaluation .

MPII [4] is a widely used benchmark for 2D human pose estimation. It includes $25K$ images collected from on-line videos. With rich diversity in backgrounds, people, clothes, lightings, and so on, 2D pose datasets like MPII significantly enhance the robustness of the trained model. Recent works [35, 54] shows that using 2D datasets contributes to a considerable performance boost compared to using 3D datasets only. To make the comparison fair, all models involved in the experiments are trained with the following two strategies:

- *Strategy 1*: only 3D data is used;
- *Strategy 2*: using additional 2D datasets in training; about 25% of data in each batch is from MPII.

5.3 Ablation Study

The effects of the two key ingredients of our method, multiple predictions (hypotheses) of the pose and the pairwise hyper bone supervision, are investigated in this section .

For a fair comparison, a new evaluation metric called Minimum Mean Per Joint Error (Min-MPJPE) is introduced to indicate the minimal MPJPE among the K predictions. In case there is a unique pose output, Min-MPJPE reduces to MPJPE.

Method	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SittingD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Zhou [59]	87.4	109.3	87.1	103.2	116.2	143.3	106.9	99.8	124.5	199.2	107.4	118.1	114.2	79.4	97.7	113.0
Tekin [45]	102.4	147.7	88.8	125.3	118.0	112.4	129.2	138.9	225.0	118.4	182.7	138.8	55.1	126.3	65.8	125.0
Zhou [57]	91.8	102.4	97.0	98.8	113.4	125.2	90.0	93.8	132.2	159.0	106.9	94.4	126.0	79.0	99.0	107.3
Sun [42]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	92.4
Mehta [29]	62.6	78.1	63.4	72.5	88.3	63.1	74.8	106.6	138.7	78.8	93.8	73.9	55.8	82.0	59.6	80.5
Pavlakos [36]	67.4	72.0	66.7	69.1	72.0	77.0	65.0	**68.3**	**83.7**	96.5	**71.7**	65.8	74.9	**59.1**	63.2	71.9
Mehta [28]	58.2	67.3	61.2	65.7	75.82	62.2	64.6	82.0	93.0	68.8	84.5	65.1	57.6	72.0	63.6	69.9
Habibie [16]	**54.0**	65.1	58.5	**62.9**	67.9	**54.0**	60.6	82.7	98.2	63.3	75.0	61.2	**50.0**	66.9	**56.5**	65.7
Sun [43]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	64.1
Ours[1](Min-MPJPE)	56.3	**61.2**	58.7	63.1	65.8	55.9	**57.6**	71.6	84.3	61.8	73.0	**58.0**	52.4	66.1	59.5	63.3
Ours[1](Fitting)	60.8	67.1	63.7	67.5	68.2	59.9	67.6	77.4	92.0	67.1	80.2	63.2	56.4	73.0	63.3	68.5
Ours[1](Regression)	57.7	62.1	**58.4**	63.5	**64.0**	56.1	60.6	68.8	86.1	**60.3**	72.1	59.7	52.2	66.3	58.2	**63.1**

Table 2: Quantitative result on Human3.6M under Protocol #1. Numbers are Mean Per Joint Position Error (MPJPE) in millimeter. Ours[1] means using *strategy 1* as stated in Section 5.2.

Multiple predictions of the pose: To study the influence of predicting multiple outputs, a network only supervised by heat maps and bone maps at the root joint is trained. As we can see from Table 1, the accuracy of a single prediction is close to the one reported by Mehta *et al.* [29], while our multiple prediction strategy leads to a significant improvement in reconstruction accuracy. The relative error decreases by more than 16%. It is observed that the multiple prediction strategy offers a relatively soft supervision, which is helpful for the network to handle ambiguities. As shown in Figure 7, the predictions with high accuracy usually come from those joints near the root joint (pelvis).

Hyper-bone Supervision: We investigate the effect of the proposed hyper bone map supervision versus the bone map supervision. Note that the element-wise product operation between the heat map and hyper bone maps is very similar to the attention mechanism [12, 18]. Each group of 3D predictions is encouraged to pay more attention to those regions with higher confidence in its associated heat map. This enables it to make a more accurate prediction. By introducing the hyper bone map supervision, the relative error decreases by more than 7% (see Table 1).

5.4 Comparison With State-of-the-art

Our method is compared favorably with those approaches using a fully convolutional framework for single person 3D pose estimation. We do not use any fully connected layers or 3D convolution operations in our convolutional pose prediction network. This greatly reduces the parameter number of our network and makes its inference fast and robust. Among all the methods that are going to be compared below, only Mehta *et al.* [27, 29] have reported real-time performance. Our method can run in real time and get comparable or better accuracy, which is a result of applying our novel bone-map representation and the corresponding loss function.

Human3.6M: The accuracy of our network predictions on Human3.6M is listed in Table 2. Our method achieves results better than the state-of-the-art approaches without including in-the-wild images with 2D key points for supervision (*Strategy 1*). Mean Per Joint Position Error (MPJPE) in millimeter is employed as our evaluation metric. Note that no information from ground truth 3D poses, such as the scale of the skeleton or the depth from the camera, is used during the evaluation. Our method improve the MPJPE by $17.4mm$ compared with the method of Mehta *et al.* [29] that adopts the joint location map representation. Another result reported in MPJPE after Procrustes alignment (PA-MPJPE) is shown in Table 3. The mixed data from MPI-INF-3DHP augmented with different backgrounds, and cloth textures provide some domain shift, but in the meantime, it also brings certain inconsistency of the key point definition. Although our training data

Method	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SittingD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Akhter [3]	199.2	177.6	161.8	197.8	176.2	186.5	195.4	167.3	160.7	173.7	177.8	181.9	176.2	198.6	192.7	181.1
Ramakrishna [38]	137.4	149.3	141.6	154.3	157.7	158.9	141.8	158.1	168.6	175.6	160.4	161.7	150.0	174.8	150.2	157.3
Zhou [58]	99.7	95.8	87.9	116.8	108.3	107.3	93.5	95.3	109.1	137.5	106.0	102.2	106.5	110.4	115.2	106.7
Moreno-Noguer [30]	66.1	61.7	84.5	73.7	65.2	67.2	60.9	67.3	103.5	74.6	92.6	69.6	71.5	78.0	73.2	74.0
Pavlakos [36]	47.5	50.5	48.3	49.3	50.7	55.2	46.1	48.0	61.1	78.1	51.1	48.3	52.9	41.5	46.4	51.9
Martinez [25]	39.5	43.2	46.4	47.0	51.0	56.0	41.4	40.6	56.5	69.4	49.2	45.0	49.5	38.0	43.1	47.7
Pavlakos [35]	**34.7**	**39.8**	41.8	**38.6**	**42.5**	47.5	**38.0**	**36.6**	**50.7**	56.8	**42.6**	**39.6**	43.9	**32.1**	**36.5**	**41.8**
ours[1](Regression)	40.0	44.2	43.1	46.6	46.3	39.4	42.3	55.0	70.4	44.7	51.8	43.5	36.8	47.5	41.2	46.2
ours[2](Regression)	36.7	41.0	**39.7**	43.0	**42.5**	**36.8**	39.0	50.5	63.8	**41.8**	48.0	40.7	**34.4**	45.5	37.7	42.7

Table 3: Comparison with previous work on Human3.6M. Protocol #1 is used. Evaluation metric is averaged PA Joint Error. Note that [35] use additional in-the-wild 2D pose dataset with ordinal depth annotations. Ours[1] (Ours[2]) means using *strategy 1* (*strategy 2*) as stated in Section 5.2.

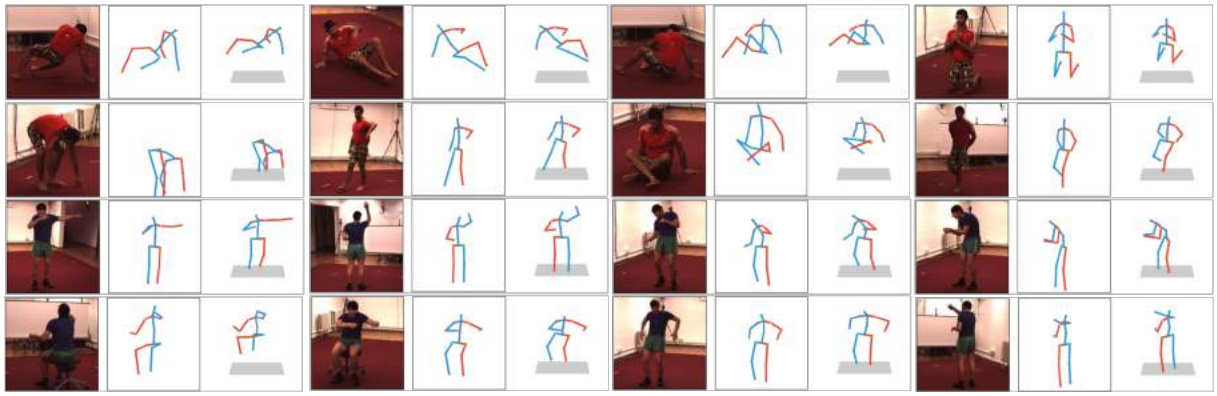

Figure 8: Ours results tested on Human3.6M. For every three columns in the figure, left: input image; middle: our 2d prediction; right: our 3d predictions. The root joint (pelvis) of all 2D predictions is relocated to the center of the image. Note that challenging examples with relatively large self-occlusions are intentionally picked and demonstrated.

Method	Tome [46]	Rogez [40]	Pavlakos [36]	Nie [33]	Tekin [44]	Zhou [55]	Martinez [25]
MPJPE	88.4	87.7	71.9	97.5	69.7	64.9	62.9
Method	Sun[42]	Mehta [28]	Mehta [27]	Yang [54]	Pavlakos [35]	ours[2](Fitting)	ours[2](Regression)
MPJPE	59.1	69.9	63.6	58.6	**56.2**	65.3	59.6

Table 4: Comparison with previous work on Human3.6M. Protocol #1 and *Strategy 2* is used. Extra 2D training data is used in all of the methods. Note that [35] use additional in-the-wild 2D pose dataset with ordinal depth annotations.

is mixed with about 25% data from MPI-INF-3DHP, our accuracy on Human3.6M is still comparable to or even better than the state-of-the-art methods. Some qualitative results are shown in Figure 8. To study the effect of adding extra 2D training to our formulation, we further train the network using a mixed training strategy (*Strategy 1*). Table 1 and Table 4 show that our method obtains accuracy that is comparable with the state-of-the-art approaches.

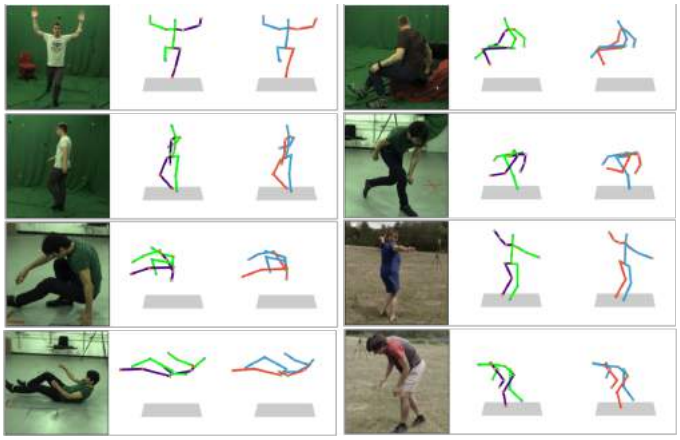

Figure 9: Qualitative results on MPI-INF-3DHP. The first column: input images; the middle column: ground truth; the last column: our predictions.

Network	Walk	Exerc.	Sit	Reach	Floor	Sport	Misc.	Total	
	PCK	PCK	PCK	PCK	PCK	PCK	PCK	PCK	AUC
Mehta [26]	86.6	75.3	74.8	73.7	52.2	82.1	77.5	75.7	39.3
Mehta [29]	88.0	81.8	78.6	77.4	59.3	82.8	81.2	79.4	41.6
Mehta [28]	83.8	75.0	77.8	77.5	55.1	80.4	72.5	75.2	37.8
Zhou [55]	85.4	71.0	60.7	71.4	37.8	70.9	74.4	69.2	32.5
habibie [16]*	-	-	-	-	-	-	-	82.9	45.4
Ours[1]	88.8	74.2	80.7	74.6	51.3	83.7	80.9	78.0	41.2
Ours[1]*	**93.0**	**85.0**	**92.0**	**89.3**	**71.9**	**94.4**	**89.2**	**88.7**	**53.1**

Table 5: PCK and AUC evaluation on the MPI-INF-3DHP dataset. Though not specially trained on this dataset, our method can achieve results better than Metha [29], which is the basic method that we aim to improve. These results demonstrate the generalization ability of our method. Both PCK and AUC metrics are the larger the better. Our pose regressor is only trained on Human3.6M dataset, so we provide results after the Procrustes Alignment (indicated by "*") for a better reference.

Most two-step methods and some approaches such as Sun *et al.* [42, 43] achieve state-of-the-art accuracy on Human3.6M, but they assume the availability of the camera intrinsics and ground-truth depth of the subject from the camera to convert their 2D poses in image coordinate space to the world coordinate space. Differently, our method do not make use of these information and thus can generate results that are more representative of the deployed system's performance.

MPI-INF-3DHP: Some results on the MPI-INF-3DHP dataset are demonstrated to analyze the generalization ability of our network in Figure 9. The accuracy of our network predictions on this dataset is listed in Table 5. Following the instruction in [26, 55], the results with the 3DPCK metric and the AUC metric are reported. Note that our network is not specifically trained for the MPI-INF-3DHP dataset, though about 25% of our training data for the ConvNet comes from it. Especially, our pose regressor is only trained with data in Human3.6M, as described in Section 4.2. To alleviate the possible influence of different scales and view angles

in two 3D datasets, we also report results after doing the Procrustes analysis for a better evaluation. Our method can achieve similar performance compared with state-of-the-art methods, indicating the generalization ability of our network.

6 CONCLUSIONS

In this paper, a robust and real-time 3D human pose estimation approach is presented. The proposed bone map representation combines 2D and 3D pose prediction into a unified framework, which enables our network to produce structure-aware and accurate pose predictions by using both global and local information. A series of experiments are conducted to evaluate our method with standard benchmarks, and it shows that the estimation accuracy of our approach is comparable to the state-of-the-art methods. More importantly, the simplicity of our network architecture leads to real-time performance.

As future work, since the pose regressor for final pose prediction only utilizes a limited set of 3D poses from the bone maps, it will be worth exploring more contextual information from the bone maps, although it may slightly raise the computational cost. It will be also interesting to extend the scalable bone map representation to address multi-person 3D pose estimation in real time.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their comments and suggestions. This work was partially supported by NSFC (61972160), Guangdong Basic and Applied Basic Research Foundation (2020A1515110997), and the Science and Technology Program of Guangzhou, China (202002030263).

ORCID

Guodong Wei, http://orcid.org/0000-0001-6975-9865
Keke Tang, http://orcid.org/0000-0003-0377-1022

REFERENCES

[1] Agarwal, A.; Triggs, B.: Recovering 3d human pose from monocular images. IEEE transactions on pattern analysis and machine intelligence, 28(1), 44–58, 2006.

[2] Akhter, I.; Black, M.J.: Pose-conditioned joint angle limits for 3d human pose reconstruction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1446–1455, 2015.

[3] Akhter, I.; Black, M.J.: Pose-conditioned joint angle limits for 3d human pose reconstruction. In Proc. IEEE Conf. on Computer Vision & Pattern Recognition, 2015.

[4] Andriluka, M.; Pishchulin, L.; Gehler, P.; Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In Proc. IEEE Conf. on Computer Vision & Pattern Recognition, 3686–3693, 2014.

[5] Andriluka, M.; Roth, S.; Schiele, B.: Monocular 3d pose estimation and tracking by detection. In Proc. IEEE Conf. on Computer Vision & Pattern Recognition, 623–630. IEEE, 2010.

[6] Bogo, F.; Kanazawa, A.; Lassner, C.; Gehler, P.; Romero, J.; Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In Proc. Euro. Conf. on Computer Vision, 561–578. Springer, 2016.

[7] Burenius, M.; Sullivan, J.; Carlsson, S.: 3d pictorial structures for multiple view articulated pose estimation. In Proc. IEEE Conf. on Computer Vision & Pattern Recognition, 3618–3625, 2013.

[8] Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of (CVPR) Computer Vision and Pattern Recognition, 1302 – 1310, 2017.

[9] Chen, C.H.; Ramanan, D.: 3d human pose estimation= 2d pose estimation+ matching. In CVPR, vol. 2, 6, 2017.

[10] Chen, L.; Wei, H.; Ferryman, J.: A survey of human motion analysis using depth imagery. Pattern Recognition Letters, 34(15), 1995–2006, 2013.

[11] Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; Chua, T.S.: Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In Proc. IEEE Conf. on Computer Vision & Pattern Recognition, 6298–6306. IEEE, 2017.

[12] Chu, X.; Yang, W.; Ouyang, W.; Ma, C.; Yuille, A.L.; Wang, X.: Multi-context attention for human pose estimation. arXiv preprint arXiv:1702.07432, 1(2), 2017.

[13] Fabbri, M.; Lanzi, F.; Calderara, S.; Alletto, S.; Cucchiara, R.: Compressed volumetric heatmaps for multi-person 3d pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7204–7213, 2020.

[14] Fan, H.; Su, H.; Guibas, L.J.: A point set generation network for 3d object reconstruction from a single image. In Proc. IEEE Conf. on Computer Vision & Pattern Recognition, vol. 2, 6, 2017.

[15] Guan, P.G.P.; Weiss, A.; Balan, A.O.; Black, M.J.: Estimating human shape and pose from a single image. In IEEE International Conference on Computer Vision, 2010.

[16] Habibie, I.; Xu, W.; Mehta, D.; Pons-Moll, G.; Theobalt, C.: In the wild human pose estimation using explicit 2d features and intermediate 3d representations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 10905–10914, 2019.

[17] He, K.; Zhang, X.; Ren, S.; Sun, J.: Deep residual learning for image recognition. In Proc. IEEE Conf. on Computer Vision & Pattern Recognition, 770–778, 2016.

[18] Hu, J.; Shen, L.; Sun, G.: Squeeze-and-excitation networks. arXiv preprint arXiv:1709.01507, 7, 2017.

[19] Ionescu, C.; Carreira, J.; Sminchisescu, C.: Iterated second-order label sensitive pooling for 3d human pose estimation. In Proc. IEEE Conf. on Computer Vision & Pattern Recognition, 1661–1668, 2014.

[20] Ionescu, C.; Papava, D.; Olaru, V.; Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE transactions on pattern analysis and machine intelligence, 36(7), 1325–1339, 2014.

[21] Kolotouros, N.; Pavlakos, G.; Black, M.; Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2020.

[22] Li, C.; Lee, G.H.: Generating multiple hypotheses for 3d human pose estimation with mixture density network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 9887–9895, 2019.

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

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

[25] Martinez, J.; Hossain, R.; Romero, J.; Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In ICCV, 2017.

[26] Mehta, D.; Rhodin, H.; Casas, D.; Fua, P.; Sotnychenko, O.; Xu, W.; Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In 3D Vision (3DV), 2017 International Conference on, 506–516. IEEE, 2017.

[27] Mehta, D.; Sotnychenko, O.; Mueller, F.; Xu, W.; Elgharib, M.; Fua, P.; Seidel, H.P.; Rhodin, H.; Pons-Moll, G.; Theobalt, C.: Xnect: Real-time multi-person 3d motion capture with a single rgb camera. ACM Transactions on Graphics (TOG), 39(4), 82–1, 2020.

[28] Mehta, D.; Sotnychenko, O.; Mueller, F.; Xu, W.; Sridhar, S.; Pons-Moll, G.; Theobalt, C.: Single-shot multi-person 3d pose estimation from monocular rgb input. In 3D Vision (3DV), 2018 Sixth International Conference on, vol. 3, 2018.

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

[30] Moreno-Noguer, F.: 3d human pose estimation from a single image via distance matrix regression. In Proc. IEEE Conf. on Computer Vision & Pattern Recognition, 1561–1570. IEEE, 2017.

[31] Mori, G.; Malik, J.: Recovering 3d human body configurations using shape contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(7), 1052–1062, 2006.

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

[33] Nie, B.X.; Wei, P.; Zhu, S.C.: Monocular 3d human pose estimation by predicting depth on joints. In IEEE International Conference on Computer Vision, 2017.

[34] Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A.A.; Tzionas, D.; Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[35] Pavlakos, G.; Zhou, X.; Daniilidis, K.: Ordinal depth supervision for 3d human pose estimation. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[36] Pavlakos, G.; Zhou, X.; Derpanis, K.G.; Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3d human pose. In Proc. IEEE Conf. on Computer Vision & Pattern Recognition, 1263–1272. IEEE, 2017.

[37] Pavlakos, G.; Zhu, L.; Zhou, X.; Daniilidis, K.: Learning to estimate 3d human pose and shape from a single color image, 2018.

[38] Ramakrishna, V.; Kanade, T.; Sheikh, Y.: Reconstructing 3d human pose from 2d image landmarks. In European conference on computer vision, 573–586. Springer, 2012.

[39] Rhodin, H.; Robertini, N.; Casas, D.; Richardt, C.; Seidel, H.P.; Theobalt, C.: General automatic human shape and motion capture using volumetric contour cues. In European Conference on Computer Vision, 509–526. Springer, 2016.

[40] Rogez, G.; Weinzaepfel, P.; Schmid, C.: Lcr-net: Localization-classification-regression for human pose. In CVPR 2017-IEEE Conference on Computer Vision & Pattern Recognition, 2017.

[41] Sarafianos, N.; Boteanu, B.; Ionescu, B.; Kakadiaris, I.A.: 3d human pose estimation: A review of the literature and analysis of covariates. Computer Vision and Image Understanding, 152, 1–20, 2016.

[42] Sun, X.; Shang, J.; Liang, S.; Wei, Y.: Compositional human pose regression. In The IEEE International Conference on Computer Vision (ICCV), vol. 2, 7, 2017.

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

[44] Tekin, B.; Marquez Neila, P.; Salzmann, M.; Fua, P.: Learning to fuse 2d and 3d image cues for monocular body pose estimation. In Proc. Int. Conf. on Computer Vision, EPFL-CONF-230311, 2017.

[45] Tekin, B.; Rozantsev, A.; Lepetit, V.; Fua, P.: Direct prediction of 3d body poses from motion compensated sequences. In Proc. IEEE Conf. on Computer Vision & Pattern Recognition, 991–1000, 2016.

[46] Tome, D.; Russell, C.; Agapito, L.: Lifting from the deep: Convolutional 3d pose estimation from a single image. Proc. IEEE Conf. on Computer Vision & Pattern Recognition, 2500–2509, 2017.

[47] Toshev, A.; Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In Proc. IEEE Conf. on Computer Vision & Pattern Recognition, 1653–1660, 2014.

[48] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I.: Attention is all you need. In Advances in Neural Information Processing Systems, 5998–6008, 2017.

[49] Wang, C.; Wang, Y.; Lin, Z.; Yuille, A.L.; Gao, W.: Robust estimation of 3d human poses from a single image. In Proc. IEEE Conf. on Computer Vision & Pattern Recognition, 2361–2368, 2014.

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

[51] Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S.: Cbam: Convolutional block attention module. In Proc. of European Conf. on Computer Vision (ECCV), 2018.

[52] Wu, J.; Xue, T.; Lim, J.J.; Tian, Y.; Tenenbaum, J.B.; Torralba, A.; Freeman, W.T.: Single image 3d interpreter network. In European Conference on Computer Vision, 365–382. Springer, 2016.

[53] Yang, W.; Li, S.; Ouyang, W.; Li, H.; Wang, X.: Learning feature pyramids for human pose estimation. In Proc. Int. Conf. on Computer Vision, vol. 2, 2017.

[54] Yang, W.; Ouyang, W.; Wang, X.; Ren, J.; Li, H.; Wang, X.: 3d human pose estimation in the wild by adversarial learning. In Proc. IEEE Conf. on Computer Vision & Pattern Recognition, vol. 1, 2018.

[55] Zhou, X.; Huang, Q.; Sun, X.; Xue, X.; Wei, Y.: Towards 3d human pose estimation in the wild: a weakly-supervised approach. In Proc. Int. Conf. on Computer Vision, 2017.

[56] Zhou, X.; Leonardos, S.; Hu, X.; Daniilidis, K.: 3d shape estimation from 2d landmarks: A convex relaxation approach. In Proc. IEEE Conf. on Computer Vision & Pattern Recognition, 4447–4455, 2015.

[57] Zhou, X.; Sun, X.; Zhang, W.; Liang, S.; Wei, Y.: Deep kinematic pose regression. In European Conference on Computer Vision, 186–201. Springer, 2016.

[58] Zhou, X.; Zhu, M.; Leonardos, S.; Daniilidis, K.: Sparse representation for 3d shape estimation: A convex relaxation approach. IEEE transactions on pattern analysis and machine intelligence, 39(8), 1648–1661, 2017.

[59] Zhou, X.; Zhu, M.; Leonardos, S.; Derpanis, K.G.; Daniilidis, K.: Sparseness meets deepness: 3d human pose estimation from monocular video. In Proceedings of the IEEE conference on computer vision and pattern recognition, 4966–4975, 2016.