# Learning Depth-aware Heatmaps for 3D Human Pose Estimation in the Wild

Zerui Chen[1,3]
zerui.chen@cripac.ia.ac.cn
Yiru Guo[5]
guoyiru0616@163.com
Yan Huang[1,3]
yhuang@nlpr.ia.ac.cn
Liang Wang[1,2,3,4]
wangliang@nlpr.ia.ac.cn

[1] Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR)

[2] Center for Excellence in Brain Science and Intelligence Technology (CEBSIT), Institute of Automation, Chinese Academy of Sciences (CASIA)

[3] University of Chinese Academy of Sciences (UCAS)

[4] Chinese Academy of Sciences Artificial Intelligence Research (CAS-AIR)

[5] School of Astronautics, Beihang University

**Abstract**

In this paper, we explore to determine 3D human pose directly from monocular image data. While current state-of-the-art approaches employ the volumetric representation to predict per voxel likelihood for each human joint, the network output is memory-intensive, making it hard to function on mobile devices. To reduce the output dimension, we intend to decompose the volumetric representation into 2D depth-aware heatmaps and joint depth estimation. We propose to learn depth-aware 2D heatmaps via associative embeddings to reconstruct the connection between the 2D joint location and its corresponding depth. Our approach achieves a good trade-off between complexity and high performance. We conduct extensive experiments on the popular benchmark Human3.6M [6] and advance the state-of-the-art accuracy for 3D human pose estimation in the wild.

## 1 Introduction

3D human pose estimation is to estimate the full body 3D pose of a human from a single monocular image, which provides comprehensive knowledge for tasks such as action recognition, autonomous driving, human-computer interaction. Due to its ill-posed nature, though significant advances have been witnessed in this task recently, it still remains an open challenge.

Existing end-to-end learning approaches attempt to localize 3D coordinates directly from a single image by addressing it as coordinate regression [8, 22], volumetric heatmaps prediction [16, 21] or other variant approaches based on 2D marginal heatmaps [14]. Compared

to these regression approaches, volumetric heatmaps directly predict per voxel likelihood for each joint. By modeling the joint located directly in the discretized 3D space, volumetric heatmaps make the most of spatial context information and achieve state-of-the-art accuracy. However, high-dimensional representation consumes too much memory and needs to take several training stages to optimize the large-scale network parameters [16]. In real practice, we often need to perform multi-person 3D pose estimation in crowded and wild scenes simultaneously, which brings great computational pressure for those approaches.

Beyond that, the largest 3D human pose dataset [6] is captured in controlled lab environments, and models trained on it cannot generalize well to in-the-wild images. The solution [20] is to perform 3D pose learning together with 2D pose data which contains abundant in-the-wild images and 2D annotations correspondingly. Benefiting from the rapid development in 2D human pose estimation, the localization of human joints in 2D space is relatively precise nowadays. Thus, we intend to establish an end-to-end pipeline which is more compatible with in-the-wild 2D pose data.

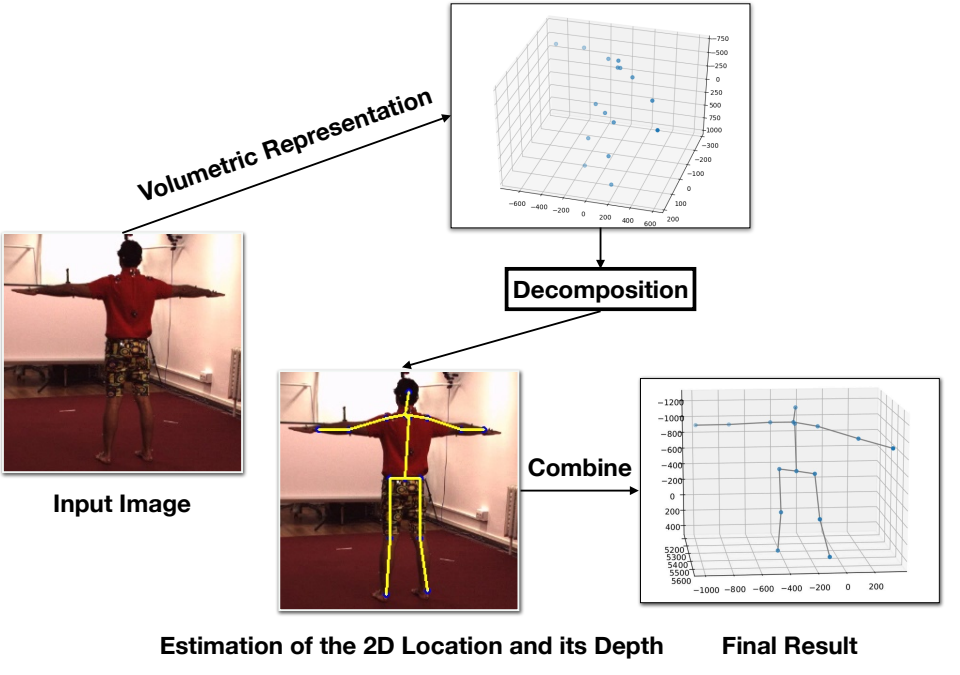


Figure 1: Illustration of our approach, which consists of two subtasks: one is the localization of human joints in 2D space, and the other is the joint depth estimation.

As illustrated in Figure 1, instead of employing the volumetric representation, we propose to untangle it as 2D marginal depth-aware heatmaps via associative embeddings and joint depth estimation. 2D marginal depth-aware heatmaps are highly compatible with in-the-wild 2D pose data and provide our pipeline with more flexible structures. Based on structural dynamics of the human body, though people can have different gestures in different scenarios, the relative locations between some paired human joints can be roughly fixed

or vary in limited magnitude. Therefore, there may exist a statistical law on the depth distribution for each human joint. To fully exploit the depth prior, we propose to learn associative embeddings and generate depth-aware heatmaps, which encode the hidden prior knowledge of joint depth.

Without sacrificing estimation performance, our approach significantly relieves memory and computational complexity with a relatively low-dimensional output, which is of great significance in real practice. The effectiveness of our approach is validated by comprehensive experiments, rigorous ablation study, and comparison with previous state-of-the-art approaches on the largest 3D benchmark *Human3.6M* [6]. Notably, our approach performs high-quality 3D human pose estimation for in-the-wild images and advances the state-of-the-art accuracy with 48.4 *mm* averaged *Joint Error*.

## 2 Related Work

Since approaches to estimating human 3D pose vary in different settings, we survey approaches that are most relevant to ours with a focus on CNN-based approaches. For a complete literature review, we refer the reader to a survey [19].

### 2.1 Joint Regression

Many CNN-based approaches cast the task of 3D human pose estimation as 3D coordinate regression. The target output is the spatial $x$, $y$, $z$ coordinates of the human joints with respect to a known root joint, commonly set as the pelvis. Li *et al*. [8] use maps for 2D joint classification to benefit their network. Tekin *et al*. [22] enforce structural constraints on the output by applying the autoencoder network. Park *et al*. [15] combine the 2D joint predictions and image features with improving the performance for the 3D joint localization. Zhou *et al*. [29] embed a kinematic model to correct abnormal regressed pose. By employing viewpoint prediction, Ghezelghieh *et al*. [4] provide the network with global joint configuration information. Sun *et al*. [20] propose a bone based representation for joints localization to enhance the connection between joint locations.

### 2.2 Volumetric Representation

The volumetric representation is to model the joint located directly in the discretized 3D space and predict per voxel likelihood for each joint independently. Compared with approaches based on joint regression with the output of low-dimensional vector of joint locations, volumetric heatmaps provide a richer output. By more adequately exploiting the power of spatial context information, approaches based on the volumetric representation achieve the state-of-the-art accuracy. By employing the volumetric representation, Pavlakos *et al*. [16] propose a coarse-to-fine learning procedure for 3D human pose estimation. In [21], with the help of the volumetric representation, integral loss further improves the estimation accuracy and advances the state-of-the-art accuracy with 49.6 *mm* averaged *Joint Error*. To relieve the heavy memory burden, Nibali *et al*. [14] replace the volumetric representation with three 2D marginal heatmaps, one for $x$-$y$ axis, one for $x$-$z$ axis and the other for $y$-$z$ axis. Zhou *et al*. [30] perform 3D human pose estimation with 2D heatmaps and depth estimation and is most related to us. However, we go further to learn associative embeddings to refine 2D heatmaps and advance estimation accuracy with a simpler structure, which is more convenient for application in real practice.

# 3 Technical Approach

The following subsections summarize our technical approach. Section 3.1 describes our proposed depth-aware heatmaps via associative embeddings and discusses their merits. Section 3.2 describes the overall structure of our network in detail.

## 3.1 Depth-aware Heatmaps via Associative Embeddings

Approaches based on the volumetric representation intend to directly model the joint located in the discretized 3D space and better preserve spatial context information. However, as it is shown in Table 1, in terms of space and time complexity, this kind of method is unfriendly to put into real practice, especially for 3D multi-person estimation on mobile devices. To reduce model complexity, we intend to decompose the volumetric representation into 2D marginal heatmaps and depth estimation.

| Complexity Comparison | Representation Dimension | Time Complexity |
|---|---|---|
| Volumetric Representation | 262144 ( $64 \times 64 \times 64$ ) | $O(N^3)$ |
| Our Baseline Method | 4097 ( $64 \times 64 + 1$ ) | $O(N^2)$ |
| Our Proposed Method | 4098 ( $64 \times 64 + 2$ ) | $O(N^2)$ |

Table 1: Comparison with our baseline method and the volumetric representation. Representation dimension is the number of parameters used to locate each joint. Time complexity in the post-processing procedure is also compared.

However, it is not a trivial task to reduce the dimension of network output without the cost of performance. Undoubtedly, the lower dimension of network output results in loss of 3D human body structural information and causes performance degradation with high probability. To validate the effectiveness of our method, we build our baseline model as in Figure 4, which directly detects the 2D joint location in heatmaps and regresses joint depth.

Based on our analysis, we consult to depth prior to each human joint. Although human gestures vary in different shapes or directions, they are always subject to human structural dynamics, and the depth for individual human joint varies in limited ranges. We can observe that human body structure has a high degree of symmetry and joint depth has a normal distribution. Therefore, we can utilize the prior distribution of joint depth to estimate the 3D joint location more accurately.

Inspired by [13] which predicts associative embeddings to group joints within the same instance, we are in an effort to encode the relative positional relationship between human joints in 2D marginal heatmaps. We propose to learn associative embeddings and cluster the depth information for each specific human joint. For a more detailed illustration of our approach, we reformulate the volumetric representation as to approximate the joint probability

for all joint locations:

$$
\begin{aligned}
&P\left(\boldsymbol{J}^{1}_{x,\,y,\,z}, \boldsymbol{J}^{2}_{x,\,y,\,z}, ..., \boldsymbol{J}^{n}_{x,\,y,\,z} \mid \theta\right) \approx \prod_{k=1}^{n} P\left(\boldsymbol{J}^{k}_{x,\,y,\,z} \mid \theta\right) \\
&= \prod_{k=1}^{n} P\left(\boldsymbol{J}^{k}_{x,\,y} \mid \boldsymbol{J}^{k}_{z},\, \theta\right) \cdot P\left(\boldsymbol{J}^{k}_{z} \mid \theta\right) \\
&= \prod_{k=1}^{n} \frac{P\left(\boldsymbol{J}^{k}_{x,\,y} \mid \boldsymbol{J}^{k}_{E},\, \theta\right) \cdot P\left(\boldsymbol{J}^{k}_{E} \mid \theta\right) \cdot P\left(\boldsymbol{J}^{k}_{z} \mid \boldsymbol{J}^{k}_{x,\,y}, \boldsymbol{J}^{k}_{E},\, \theta\right)}{P\left(\boldsymbol{J}^{k}_{z} \mid \theta\right) \cdot P\left(\boldsymbol{J}^{k}_{E} \mid \boldsymbol{J}^{k}_{x,\,y,\,z},\, \theta\right)} P\left(\boldsymbol{J}^{k}_{z} \mid \theta\right)
\end{aligned}
\tag{1}
$$

Where we predict the location of each human joint independently and consult to associative embeddings to encode prior depth knowledge of joint depth. $\boldsymbol{J}^{k}_{x,\,y,\,z}$ is the location of the $k_{th}$ joint, a 3D vector consisting of concatenation of the $x,\ y,\ z$ coordinates. $\theta$ represents network parameters. $\boldsymbol{J}^{k}_{E}$ is the associative embedding for $k_{th}$ joint. $P(\boldsymbol{J}^{k}_{x,\,y} \mid \boldsymbol{J}^{k}_{E},\, \theta)$ and $P\left(\boldsymbol{J}^{k}_{z} \mid \theta\right)$ are the depth-aware 2D marginal heatmap and depth distribution for $k_{th}$ joint respectively.

Since the associative embedding preserves the depth prior knowledge, its distribution should be roughly the same as the distribution of the joint depth. We can assume that:

$$
\begin{aligned}
&P\left(\boldsymbol{J}^{k}_{E} \mid \theta\right) \approx P\left(\boldsymbol{J}^{k}_{z} \mid \theta\right) \\
&P\left(\boldsymbol{J}^{k}_{z} \mid \boldsymbol{J}^{k}_{x,\,y}, \boldsymbol{J}^{k}_{E},\, \theta\right) \approx P\left(\boldsymbol{J}^{k}_{E} \mid \boldsymbol{J}^{k}_{x,\,y,\,z},\, \theta\right)
\end{aligned}
\tag{2}
$$

Based on our assumptions, the formulation in Eq. 2 can be simplified and we can obtain our final formulation:

$$
\begin{aligned}
&\prod_{k=1}^{n} \frac{P\left(\boldsymbol{J}^{k}_{x,\,y} \mid \boldsymbol{J}^{k}_{E},\, \theta\right) \cdot P\left(\boldsymbol{J}^{k}_{E} \mid \theta\right) \cdot P\left(\boldsymbol{J}^{k}_{z} \mid \boldsymbol{J}^{k}_{x,\,y}, \boldsymbol{J}^{k}_{E},\, \theta\right)}{P\left(\boldsymbol{J}^{k}_{z} \mid \theta\right) \cdot P\left(\boldsymbol{J}^{k}_{E} \mid \boldsymbol{J}^{k}_{x,\,y,\,z},\, \theta\right)} P\left(\boldsymbol{J}^{k}_{z} \mid \theta\right) \\
&\approx \prod_{k=1}^{n} P\left(\boldsymbol{J}^{k}_{x,\,y} \mid \boldsymbol{J}^{k}_{E},\, \theta\right) \cdot P\left(\boldsymbol{J}^{k}_{z} \mid \theta\right)
\end{aligned}
\tag{3}
$$

Our approach embeds the prior knowledge in 2D heat maps as shown in Eq. 3. In Figure 2, compared to original 2D heatmaps, depth-aware heatmaps have more obvious hot zones and discriminate different parts of the human body more explicitly. Our approach significantly improves estimation accuracy in 2D space and leads to overall performance improvement.

## 3.2 Network Architecture

We present our overall architecture as illustrated in Figure 3. Given an input image, we use a pretrained ResNet-50 network [9] with the U-shape network to extract feature maps. The final feature map size is 1/4 of the input image. On every stage of the U-shape network, we apply pyramid pooling [26] to produce scale-aware features. More specifically, We perform average pyramid pooling with bin sizes of $1 \times 1$, $2 \times 2$, $4 \times 4$, $6 \times 6$. By using the 4-level pyramid, pooling kernels cover the whole, half of, and small portions of the image, which is beneficial for joint depth inference.

Followed by feeding the scale-aware features to our network, we estimate the joint depth and its associative embedding for each human joint. On the basis of scale-aware feature maps, we generate depth-aware heatmaps via associative embeddings which encode the prior

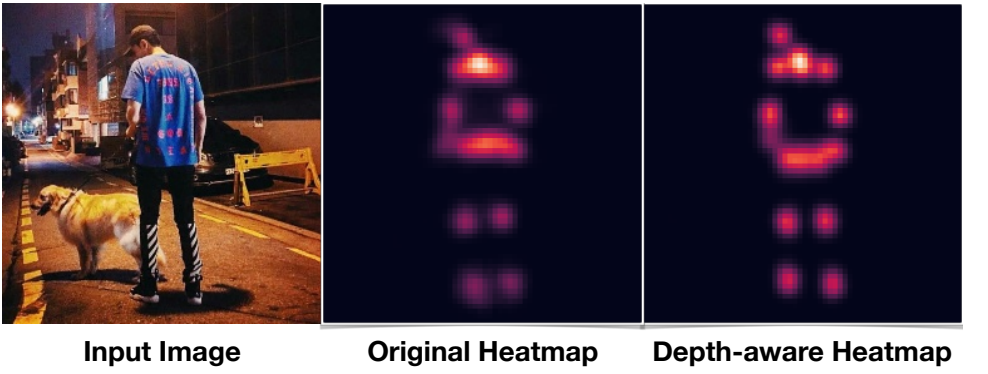


Figure 2: Visualization and Comparison between our depth-aware heatmaps and original heatmaps.

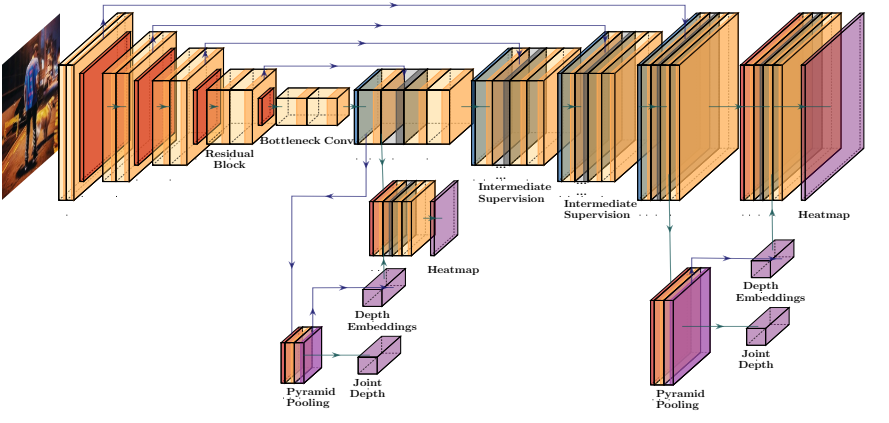


Figure 3: Overview of our network architecture. Given a monocular image, we employ ResNet-50 as our backbone network. Then feature maps are fused in U-shape to narrow the semantic gap. On every stage of the U-shape network, we produce scale-aware features and then generate depth-aware heatmaps.

depth knowledge for each specific joint and reconstruct the relationship between the 2D joint location and its corresponding depth. Our network provides abundant scale-aware and depth-aware information for efficient 3D human pose estimation. The learning procedure is in an end-to-end fashion.

Following the standard practice in 2D human pose estimation, we apply supervision on intermediate feature maps of sizes $8 \times 8$, $16 \times 16$, $32 \times 32$, respectively. These intermediate feature maps are generated by the U-shape network, and the final heatmap is $64 \times 64$ in size.

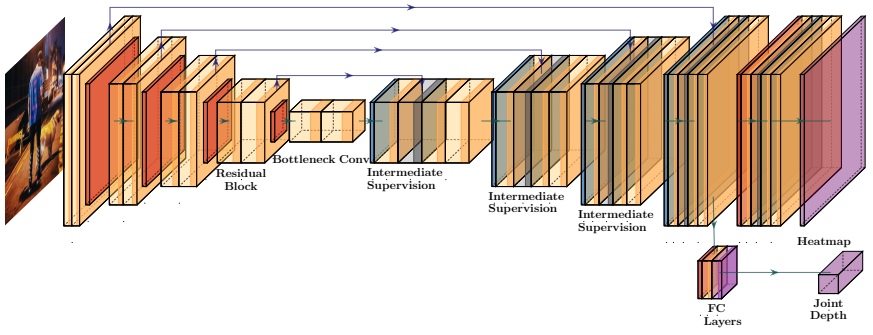


Figure 4: Overview of our baseline architecture. Given a monocular image, we employ ResNet and the U-shape neck network to extract image features. We directly regress joint depth and detect the 2D joint location.

# 4 Experiments

Our experimental evaluation focuses on 3D human pose estimation in the wild. We present an extensive evaluation of our approach to 3D human pose benchmarks. *Human3.6M* [6] is the current largest benchmark for 3D human pose estimation, consisting of 3.6 million video frames recorded from 4 different cameras and corresponding 3D human poses. It contains video of 11 subjects including a variety of actions, such as 'Walking', 'Sitting' and 'Phoning'.

*MPII* [1] is a commonly used benchmark for 2D human pose estimation. It contains about *25k* images and *40k* annotated 2D poses. The images were collected from YouTube videos covering daily human activities.

## 4.1 Comprehensive Evaluation Metric

*Protocol 1*: Six subjects (S1, S5, S6, S7, S8, S9) are used in training. Evaluation is performed on every 64th frame of Subject 11's videos. It is used in [2, 12, 18, 25, 28]. *PA Joint Error* is used for evaluation.

*Protocol 2*: Five subjects (S1, S5, S6, S7, S8) are used in training. Evaluation is performed on every 64th frame of (S9, S11). It is used in [2, 3, 7, 11, 12, 16, 23, 24, 27, 28, 29]. *Joint Error* is used for evaluation.

## 4.2 Implementation Details

We crop bounding boxes centered on the target person by using the groundtruth annotations and fill a training batch with 50% data from extra 2D training data and 50% data from *Human3.6M*. We apply *L2* loss to supervise the prediction of depth-aware heatmaps and joint depth. Intermediate supervision is employed on every stage of the U-shape network.

For training, we follow the previous practice to use $256 \times 256$ as input resolution. The resolution of each heatmap is $64 \times 64$. We use Adam optimizer with the learning rate of 1.6e-4. The batch size is 80. Training examples are dynamically augmented via horizontal

flipping and random color changes. The weight decay is set to 1e-4. The training procedure runs for 180 epochs, including 3 warm-up epochs.

## 4.3 Quantitative Results

| Method | Direct | Dicuss | Eat | Greet | Phone | Pose | Purch. | Sit | SitD | Smoke | Photo | Wait | Walk | WalkD | WalkT | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Chen [3] | 89.9 | 97.6 | 90.0 | 107.9 | 107.3 | 93.6 | 136.1 | 133.1 | 240.1 | 106.7 | 139.2 | 106.2 | 87.0 | 114.1 | 90.6 | 114.2 |
| Tome [22] | 65.0 | 73.5 | 76.8 | 86.4 | 86.3 | 68.9 | 74.8 | 110.2 | 173.9 | 85.0 | 110.7 | 85.8 | 71.4 | 86.3 | 73.1 | 88.4 |
| Moreno [13] | 69.5 | 80.2 | 78.2 | 87.0 | 100.8 | 76.0 | 69.7 | 104.7 | 113.9 | 89.7 | 102.7 | 98.5 | 79.2 | 82.4 | 77.2 | 87.3 |
| Zhou [28] | 68.7 | 74.8 | 67.8 | 76.4 | 76.3 | 84.0 | 70.2 | 88.0 | 113.8 | 78.0 | 98.4 | 90.1 | 62.6 | 75.1 | 73.6 | 79.9 |
| Jahangiri [9] | 74.4 | 66.7 | 67.9 | 75.2 | 77.3 | 70.6 | 64.5 | 95.6 | 127.3 | 79.6 | 79.1 | 73.4 | 67.4 | 71.8 | 72.8 | 77.6 |
| Mehta [14] | 57.5 | 68.6 | 59.6 | 67.3 | 78.1 | 56.9 | 69.1 | 98.0 | 117.5 | 69.5 | 82.4 | 68.0 | 55.3 | 76.5 | 61.4 | 72.9 |
| Pavlakos [16] | 58.6 | 64.6 | 63.7 | 62.4 | 66.9 | 57.7 | 62.5 | 76.8 | 103.5 | 65.7 | 70.7 | 61.6 | 56.4 | 69.0 | 59.5 | 66.9 |
| Zhou [30] | 54.8 | 60.7 | 58.2 | 71.4 | 62.0 | 65.5 | 53.8 | 55.6 | 75.2 | 111.6 | 64.2 | 66.0 | 51.4 | 63.2 | 55.3 | 64.9 |
| Martinez [10] | 49.2 | 51.6 | 47.6 | 50.5 | 51.8 | 48.5 | 51.7 | 61.5 | 94.6 | 62.3 | 78.4 | 59.1 | 65.1 | 49.5 | 52.4 | 62.9 |
| Sun [20] | 52.8 | 54.8 | 54.2 | 54.3 | 61.8 | 53.1 | 53.6 | 71.7 | 86.7 | 61.5 | 67.2 | 53.4 | 47.1 | 61.6 | 53.4 | 59.1 |
| Pavlakos [17] | 48.5 | 54.4 | 54.4 | 52.0 | 59.4 | 49.9 | 52.9 | 65.8 | 71.1 | 56.6 | 65.3 | 52.9 | 60.9 | 44.7 | 47.8 | 56.2 |
| Luvizon [8] | 49.2 | 51.6 | 47.6 | 50.5 | 51.8 | 48.5 | 51.7 | 61.5 | 70.9 | 53.7 | 60.3 | 48.9 | 44.4 | 57.9 | 48.9 | 53.2 |
| Sun [21] | 47.5 | **47.7** | 49.5 | 50.2 | 51.4 | 43.8 | **46.4** | 58.9 | 65.7 | 49.4 | **55.8** | 47.8 | **38.9** | **49.0** | **43.8** | 49.6 |
| Ours | **45.3** | 49.8 | **46.1** | **49.6** | **48.2** | **41.7** | 47.4 | **53.1** | **55.2** | **48.0** | 57.7 | **45.6** | 40.8 | 52.4 | 45.2 | **48.4** |

Table 2: Comparison with previous work on *Human3.6M*. *Protocol 2* is used. Extra 2D data is used in all these methods.

Table 2 summarizes quantitative results of our approach under *Protocol 2*. Our approach outperforms the state-of-the-art accuracy by 1.2 *mm* and achieves leading results in 9 of 15 action categories. Notably, on hard actions for previous state-of-the-art methods such as 'Sitting Down', our architecture outperforms the current best accuracy [21] by more than 10 *mm*, a relative 15.7% performance improvement. A 9.8% accuracy improvement can also be observed on the 'Sitting' action.

| Method | Yasin [25] | Rogez [12] | Chen [3] | Moreno [13] | Sun [20] | Sun [21] | Ours |
|---|---|---|---|---|---|---|---|
| *PA Joint Error* (*mm*) | 108.3 | 88.1 | 82.7 | 76.5 | 48.3 | 40.6 | 33.7 |

Table 3: Comparison with previous work on *Human3.6M*. *Protocol 1* is used. Extra 2D training data is used in all these methods.

Table 3 summarizes quantitative results of our approach under *Protocol 1*. The advantage of our approach is significantly obvious under this setting. We advance the state-of-the-art accuracy by 17%, and it shows that our approach can achieve more robust estimation results. Benefitting from the exact 2D joint location obtained from depth-aware heatmaps, our estimation results match its corresponding groundtruth with high confidence and significantly reduces *Joint Error* after a certain rigid transformation.

## 4.4 Ablation Study

In order to validate the effectiveness of depth-aware heat maps and scale-aware features generated by pyramid pooling, we conduct various and rigorous ablation experiments. We build four baseline models.

**3D/wo DH:** We do not employ depth-aware heatmaps in this baseline. It only uses 3D data to train the network. In-the-wild 2D pose data is not used.

**3D/wo SF:** We do not adopt pyramid pooling to generate scale-aware features in this baseline. We only use 3D data to train the network. In-the-wild 2D pose data is not used.

**3D/wo DH, SF:** Neither Depth-aware heatmaps nor scale-aware features are employed. We only use 3D data to train the network. In-the-wild 2D pose data is not used.

**3D+2D/wo DH, SF:** Neither Depth-aware heatmaps nor scale-aware features are employed. We use 3D data together with in-the-wild 2D pose data to train the network.

Our proposed approach is denoted as **3D+2D/w DH, SF**.

| Evaluation Metric | 3D/wo DH | 3D/wo SF | 3D/wo DH, SF | 3D+2D/wo DH, SF | 3D+2D/w DH, SF |
|---|---|---|---|---|---|
| *Protocol 1* | 38.6 | 38.4 | 38.9 | 35.0 | 33.7 |
| *Protocol 2* | 53.7 | 53.1 | 54.3 | 52.2 | 48.4 |

Table 4: Results of our ablation experiments. The numbers are mean Euclidean distance (*mm*) between the groundtruth 3D joints and our predictions.

As shown in Table 4, we can draw several conclusions about this task. By comparing **3D/wo DH, SF** and **3D+2D/wo DH, SF**, we can observe that extra 2D pose data has a significant impact on the performance of the model. Not limited to improving the generalization ability, extra 2D pose data provides a proper regularization for heatmaps training, leading to more robust estimation accuracy under both protocols.

Depth-aware heatmaps (DH) intends to establish the relationship between the 2D joint location and its corresponding joint depth. In comparison with **3D/wo SF, DH**, **3D/wo SF** achieves a 1.2 *mm* accuracy improvement under the *Protocol 2*. Since extra 2D data plays a more important role in regularization, it still leads to a 0.5 *mm* improvement after a rigid alignment.

Scale-aware features (SF) generated by pyramid pooling are designed to provide fruitful semantic context information for joint depth inference. We perform average pyramid pooling with various bin sizes, and our model can attend to the whole or even small portions of the image. This technique is effective and advances estimation accuracy by 0.3 *mm* and 0.6 *mm* under the *Protocol 1* and *Protocol 2*, respectively.

## 4.5 Qualitative Results

Models only trained on *Human3.6M* cannot generalize well to the scene in the wild, and there are various reasons. First, *Human3.6M* dataset is collected in limited indoor scenes, and thus the image background is quite monotonous and lacks variations. Second, human pose diversity is also limited in *Human3.6M* dataset, making it hard to estimate 3D human pose in the 'real' scene. With the introduction of depth-aware heatmaps and scale-aware features, our model can better generalize to the scene in the wild. The qualitative result is shown in Figure 5. Since there are no 3D annotations in in-the-wild 2D pose data, we present

qualitative results on it. As we can see in Figure 5, our proposed method can generalize well to in-the-wild images and achieve robust estimation results.

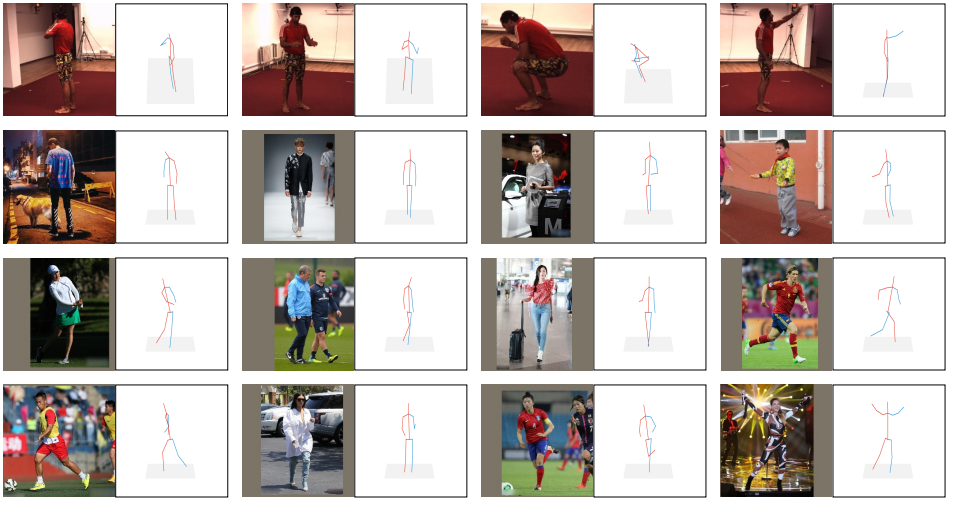

Figure 5: Examples of 3D human pose estimation for *Human3.6M* (top row) and in-the-wild 2D human pose data (middle and bottom rows). We fill some images with mean pixel values to ensure a proper aspect ratio.

## 5 Conclusion and Future Work

In this work, we propose an end-to-end pipeline for 3D human pose estimation. Compared to previous state-of-art approaches based on the volumetric representation, our approach has a much lower output dimension and provides our network with more flexible structures, which is of great significance in real practice. With the introduction of depth-aware heatmaps via their associative embeddings based on scale-aware features, we exploit the prior knowledge of joint depth to efficiently reconstruct the connection between the 2D joint location and its corresponding joint depth. Our approach advances the state-of-the-art accuracy for 3D human pose estimation with averaged *Joint Error* of 48.4 *mm* and presents us with excellent qualitative results for in-the-wild images.

In the future, we plan to explore more weakly-supervised or unsupervised methods to conduct efficient 3D human pose estimation in the wild, especially for multi-person 3D pose estimation. We hope our work can inspire more research on efficient 3D human pose estimation in the wild.

## Acknowledgements

This work is jointly supported by National Key Research and Development Program of China (2016YFB1001000), National Natural Science Foundation of China (61525306, 61633021, 61721004, 61420106015, 61806194), Capital Science and Technology Leading Talent Training Project(Z181100006318030), Beijing Science and Technology Project (Z181100008918010), and CAS-AIR.

# References

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

[2] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation= 2d pose estimation+ matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[3] Hao-Shu Fang, Yuanlu Xu, Wenguan Wang, Xiaobai Liu, and Song-Chun Zhu. Learning pose grammar to encode human body configuration for 3d pose estimation. In *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[4] Mona Fathollahi Ghezelghieh, Rangachar Kasturi, and Sudeep Sarkar. Learning camera viewpoint using cnn to improve 3d body pose estimation. In *International Conference on 3D Vision (3DV)*, 2016.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

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

[7] Ehsan Jahangiri and Alan L Yuille. Generating multiple diverse hypotheses for human 3d pose consistent with 2d joint detections. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[8] Sijin Li and Antoni B Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *Asian Conference on Computer Vision (ACCV)*, 2014.

[9] Diogo C Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[10] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[11] Dushyant Mehta, Helge Rhodin, Dan Casas, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation using transfer learning and improved cnn supervision. *In International Conference on 3D Vision (3DV)*, 2016.

[12] Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[13] Alejandro Newell, Deng Jia, and Zhiao Huang. Associative embedding: End-to-end learning for joint detection and grouping. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[14] Aiden Nibali, Zhen He, Stuart Morgan, and Luke Prendergast. 3d human pose estimation with 2d marginal heatmaps. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019.

[15] Sungheon Park, Jihye Hwang, and Nojun Kwak. 3d human pose estimation using convolutional neural networks with 2d pose information. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[16] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[17] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[18] Grégory Rogez and Cordelia Schmid. Mocap-guided data augmentation for 3d pose estimation in the wild. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[19] Nikolaos Sarafianos, Bogdan Boteanu, Bogdan Ionescu, and Ioannis A Kakadiaris. 3d human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding (CVIU)*, 2016.

[20] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[21] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[22] Bugra Tekin, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. Structured prediction of 3d human pose with deep neural networks. *Proceedings of the British Machine Vision Conference (BMVC)*, 2016.

[23] Bugra Tekin, Artem Rozantsev, Vincent Lepetit, and Pascal Fua. Direct prediction of 3d body poses from motion compensated sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[24] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[25] Hashim Yasin, Umar Iqbal, Bjorn Kruger, Andreas Weber, and Juergen Gall. A dual-source approach for 3d pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[26] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[27] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[28] Xiaowei Zhou, Menglong Zhu, Georgios Pavlakos, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. Monocap: Monocular human motion capture using a cnn coupled with a geometric prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018.

[29] Xingyi Zhou, Xiao Sun, Wei Zhang, Shuang Liang, and Yichen Wei. Deep kinematic pose regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[30] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.