MEBOW: Monocular Estimation of Body Orientation In the Wild

Chenyan Wu[1,2*]   Yukun Chen[1]   Jiajia Luo[2]   Che-Chun Su[2]   Anuja Dawane[2]

Bikramjot Hanzra[2]   Zhuo Deng[2]   Bilan Liu[2]   James Z. Wang[1]   Cheng-hao Kuo[2]

[1]The Pennsylvania State University, University Park   [2]Amazon Lab126

{czw390,yzc147,jwang}@psu.edu

{lujiajia,ccsu,adawane,hanzrabh,zhuod,liubila,chkuo}@amazon.com

Abstract

Body orientation estimation provides crucial visual cues in many applications, including robotics and autonomous driving. It is particularly desirable when 3-D pose estimation is difficult to infer due to poor image resolution, occlusion, or indistinguishable body parts. We present COCO-MEBOW (Monocular Estimation of Body Orientation in the Wild), a new large-scale dataset for orientation estimation from a single in-the-wild image. The body-orientation labels for around 130K human bodies within 55K images from the COCO dataset have been collected using an efficient and high-precision annotation pipeline. We also validated the benefits of the dataset. First, we show that our dataset can substantially improve the performance and the robustness of a human body orientation estimation model, the development of which was previously limited by the scale and diversity of the available training data. Additionally, we present a novel triple-source solution for 3-D human pose estimation, where 3-D pose labels, 2-D pose labels, and our body-orientation labels are all used in joint training. Our model significantly outperforms state-of-the-art dual-source solutions for monocular 3-D human pose estimation, where training only uses 3-D pose labels and 2-D pose labels. This substantiates an important advantage of MEBOW for 3-D human pose estimation, which is particularly appealing because the per-instance labeling cost for body orientations is far less than that for 3-D poses. The work demonstrates high potential of MEBOW in addressing real-world challenges involving understanding human behaviors. Further information of this work is available at https://chenyanwu.github.io/MEBOW/.

1. Introduction

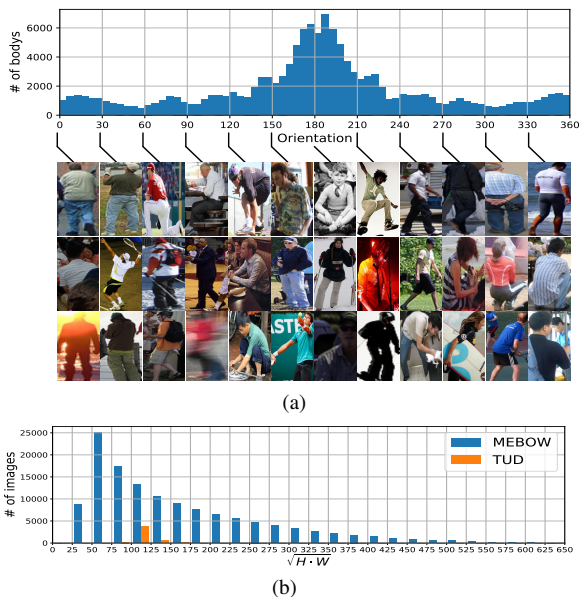*Human body orientation estimation* (HBOE) aims at es-

---

Figure 1. Overview of the MEBOW dataset. (a) Distribution of the body orientation labels in the dataset and examples. (b) Comparison of the distribution of the captured human body instance resolution for our dataset and that for the TUD dataset [6]. The x-axis represents ($\sqrt{W \times H}$), where W and H are the width and height of the human body instance bounding box in pixels, respectively.

timating the orientation of a person with respect to the camera point of view. It is important for a number of industrial applications, *e.g.*, robots interacting with people and autonomous driving vehicles cruising through crowded urban areas. Given a predicted 3-D human pose, commonly in the form of a skeleton with dozens of joints, the body orientation can be inferred. Hence, one may argue that HBOE is a simpler task compared with 3-D human pose estimation and directly solvable using pose estimation models. Nonetheless, HBOE warrants to be tackled as a standalone problem for three reasons. First, 3-D pose may be difficult to infer due to poor image resolution, occlusion, or indistinguish-

able body parts, all of which are prevalent in in-the-wild images. Second, under certain scenarios, the orientation of the body is already sufficient to be used as the cue for downstream prediction or planning tasks. Third, much reduced computational cost for body orientation model compared to a 3-D model makes it more appealing for on-device deployment. Moreover, body orientation estimation and 3-D pose estimation may be complementary in addressing real-world challenges involving understanding human behaviors.

HBOE has been studied in recent years [6, 8, 10, 14, 18, 19, 27, 33, 45, 53, 54]. A primary bottleneck, however, is a lack of a large-scale, high-precision, diverse-background dataset. Previously, the TUD dataset [6] has been the most widely used dataset for HBOE. But it only has about $5,000$ images, and the orientation labels are of low precision because they are quantized into eight bins/classes. Hara *et al.* [18] relabeled the TUD dataset with continuous orientation labels, but the scale limitation is left unaddressed. And we verify experimentally that the model trained on it generalizes much worse to in-the-wild images compared with the much larger dataset we present here. Because body orientation could be inferred from a 3-D pose label (in the form of a list of 3-D coordinates for predefined joints), 3-D human pose datasets, *e.g.*, Human3.6M, could be used to train body orientation estimation models after necessary preprocessing. However, those datasets are commonly only recorded indoors (due to the constraint of motion capture systems), with a clean background, bearing little occlusion problems, and for a limited number of human subjects. All of these limitations make it less likely for the body orientation models developed on existing 3-D pose datasets to generalize well to images captured in the wild, in which various occlusion, lighting conditions, and poses could arise. Given the enormous success of large-scale datasets in advancing vision research, such as ImageNet [13] for image classification, KITTI [15] for optical flow, and COCO [26] for object recognition and instance segmentation among many others, we believe the creation of a large-scale, high-precision dataset is urgent to the development of HBOE models, particularly those data-hungry deep learning-based ones.

In this paper, we present the COCO-MEBOW (Monocular Estimation of Body Orientation in the Wild) dataset, which consists of high-precision body orientation labels for 130K human instances within 55K images from the COCO dataset [26]. Our dataset uses 72 bins to partition the $360°$, with each bin covering only $5°$, which is much more fine-grained than all previous datasets while within the human cognition limit. The distributions of the collected orientation labels and some example cropped images of human bodies in our dataset are shown in Fig. 1(a). Details and the creation process will be introduced in Sec. 3.2. For brevity, we will call our dataset MEBOW in the rest of this paper.

To demonstrate the value of our dataset, we conducted two sets of experiments. The *first* set of experiments focused on HBOE itself. We first present a strong but simple baseline model for HBOE which is able to outperform previous state-of-the-art models [53] on the TUD dataset (with continuous orientation labels). We then compare the performance of our baseline model under four settings: *training on TUD and evaluating on MEBOW, training on MEBOW and evaluating on TUD, training on TUD and evaluating on TUD, training on MEBOW and evaluating on MEBOW.* We observe that the model trained on MEBOW generalizes well to TUD but not vice versa.

The *second* set of experiments focused on demonstrating the feasibility of boosting estimation performance through using our dataset as an additional, relative low-cost source of supervision. Our model is based on existing work on weakly-supervised 3-D human pose model using both 2-D pose dataset and 3-D pose dataset as the source of supervision. And the core of our model is a novel *orientation loss* which enables us to leverage the body orientation dataset as an additional source of supervision. We demonstrate in Sec. 4.2 that our triple-source weakly-supervised learning approach could bring significant performance gains over the baseline dual-source weakly-supervised learning approach. This shows that our dataset could be useful for not only HBOE but also other vision tasks, among which the gain in 3-D pose estimation is demonstrated in this paper.

Our **main contributions** are summarized as follows.

1. We present MEBOW, a large-scale high-precision human body orientation dataset.

2. We established a simple baseline model for HBOE, which, when trained with MEBOW, is shown to significantly outperform state-of-the-art models trained on existing dataset.

3. We developed the first triple-source solution for 3-D human pose estimation using our dataset as one of the three supervision sources, and it significantly outperforms a state-of-the-art dual-source solution for 3-D human pose estimation. This not only further demonstrates the usefulness of our dataset but also points out and validates a new direction of improving 3-D human pose estimation by using significantly lower-cost labels (*i.e.*, body orientation).

2. Related Work

Human body orientation datasets. The TUD multi-view pedestrians dataset [6] is the most widely used dataset for benchmarking HBOE models. Most recent HBOE algorithms, *e.g.*, [6, 18, 19, 53], use it for training and evaluation. This dataset consists of $5,228$ images captured outdoors, each containing one or more pedestrians, each of which is labeled with a bounding box and a body orientation. The body orientation labels only have eight bins,

i.e., {*front*, *back*, *left*, *right*, *diagonal front*, *diagonal back*, *diagonal left*, *diagonal right*}. This labeling is rather coarse-grained, and many of the images are gray-scale images. Later work [18] enhances the TUD dataset by providing continuous orientation labels, each of which is averaged from the orientation labels collected from five different labelers. There are also some other less used datasets for HBOE. Their limitations, however, make them only suitable for HBOE under highly constrained settings but not for in-the-wild applications. For example, the 3DPes dataset [7] ($1,012$ images) and CASIA gait dataset [41] ($19,139$ frames of videos capturing 20 subjects) have been used in [53] and [42, 27], respectively. And their body orientation labels are 8-bin based and 6-bin based, respectively, which are also coarse-grained. Moreover, the human bodies in the images of these two datasets are all walking pedestrians captured from a downward viewpoint by one or a few fixed outdoor surveillance cameras. The MCG-RGBD datasets [28] has a wider diversity of poses and provides depth maps in addition to the RGB images. But all its images were captured indoors and from only 11 subjects. Because human orientation can be computed given a full 3-D pose skeleton, we can convert a human 3-D pose dataset, *e.g.*, the Human3.6M dataset [20], to a body orientation dataset for HBOE research. However, due to the constraint of the motion capture system, those 3-D pose datasets often only cover indoor scenes and are sampled frames of videos for only a few subjects. These constraints make them not as rich as our MEBOW dataset, which is based on COCO [26], in both contextual information and the variety of background. The size of the Human3.6M dataset [20] (10K frames) is also much smaller than MEBOW (130K).

Human body estimation algorithms. Limited by the relative small size and the coarse-grained orientation label (either 8-bin based or 6-bin based) of existing datasets discussed above, approaches based on feature engineering and traditional classifiers [6, 45, 14, 33, 10, 54, 8], *e.g.*, SVM, have been favored for HBOE. Deep learning-based methods [42, 12] also treat HBOE as a classification problem. For example, the method in [42] uses a 14-layer classification network to predict which bin out of the eight different bins represents the orientation given an input; the method in [12] uses a 4-layer neural network as the classification network. These methods all use simple network architecture due to the small size of the available datasets for training. And the obtained model only works for certain highly constrained environment similar to that was used for collecting training images. Given the continuous orientation label provided by [18] for the TUD dataset, some recent work [18, 19, 53] attempted to address more fine-grained body orientation prediction. Most notably, Yu *et al.* [53] utilizes the key-points detection by another 2-D pose model as an additional cue for continuous orientation prediction.

Still, deep learning-based methods are held back by the lack of a large-scale HBOE dataset. Direct prediction of body orientation from an image is valid because not only labeling a training dataset is simpler but also better performance could be achieved by directly addressing the orientation estimation problem. As a supporting evidence, [16] shows that a CNN and Fisher encoding-based method taking in features extracted from 2-D images outperforms state-of-the-art methods based on 3-D information (*e.g.*, 3-D CAD models or 3-D landmarks) for multiple object orientation estimation problems.

3-D pose estimation. The lack of large training data covering diverse settings is a major problem for robust 3-D pose estimation. Efforts [52, 30, 43, 55, 49, 48] have been made to address this by using additional source of supervision, mainly 2-D pose dataset (*e.g.*, MPII [5]). The general idea is to design some novel loss for the data with weak labels (2-D pose) to penalize incorrect 3-D pose prediction on those additional data with much more diverse human subjects and background variations so that the learnt model could better generalize to those data. Our work shows a new direction following this line of research, which is to use our large-scale, high-precision, cost-effective body orientation dataset as a new source of weak supervision. Some other ideas complementary to the above idea for improving 3-D pose estimation include: (1) enforcing extra prior knowledge such as a parameterized 3-D human mesh model [17, 24, 9, 23, 22, 35, 38], the ordinal depth [36], and temporal information (such as adjacent frame consistency) [25, 39]; and (2) leveraging images simultaneously captured from different views [40, 21], mainly for indoor dataset collected in a highly constrained environment (*e.g.*, Human3.6M).

3. The Method

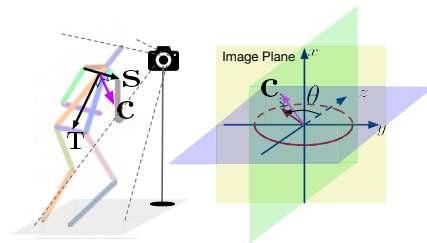## 3.1. Definition of Body Orientation

Figure 2. Definition of body orientation.

Previous datasets including TUD all assume that the human body orientation is self-explanatory from the image, which is adequate for small dataset with a consistent camera point of view. For large dataset of in-the-wild images containing all kinds of human poses and camera points of view, a formal definition of the human orientation is necessary for both annotation and modeling. As illustrated in

Fig. 2, without loss of generality, we define the human orientation $\theta \in [0°, 360°)$ as the angle between the projection vector of the chest facing direction (**C**) onto the y-z plane and the direction of the axis z, where the x, y, z vectors are defined by the image plane and the orientation of the camera. Given a 3-D human pose, the chest facing direction **C** can be computed by $\mathbf{C} = \mathbf{T} \times \mathbf{S}$, where **S** is the shoulder direction defined by the vector from the right shoulder to the left one, and **T** is the torso direction defined by the vector from the midpoint of the left- and right-shoulder joints to the midpoint of the left- and right-hip joints.

3.2. MEBOW Dataset Creation

We choose the COCO dataset [26] as the source of images for orientation labeling for the following reasons. First, the COCO dataset has rich contextual information. And the diversity of human instances captured within the COCO dataset in terms of poses, lighting condition, occlusion types, and background makes it suitable for developing and evaluating models for body orientation estimation in the wild. Secondly, the COCO dataset already has bounding box labels for human instances, making it easier for body orientation labeling. To make our dataset large scale, after neglecting ambiguous human instances, we labeled all suitable $133,380$ human instances within the total $540,007$ images, out of which $51,836$ images (associated with $127,844$ human instances) are used for training and $2,171$ images (associated with $5,536$ human instances) for testing. To our knowledge, MEBOW is the largest HBOE dataset. The number of labeled human instances in our dataset is about 27 times that of TUD. To make our dataset of high precision, we choose a 72-bin annotation scheme, which not only is much more fine-grained than former 8-bin or 6-bin annotation used by other HBOE datasets, but also accounts for the cognitive limits of human labelers and the variance of labels between different labelers. Fig. 1(a) shows the distribution of our orientation labels, along with some example human instances. It can be seen that our dataset covers all possible body orientation, with a Gaussian like peak around $180°$, which is natural because photos with humans tend to capture the main person from the front. Another advantage of our dataset is that the image resolution of the labeled human instances is much more diverse than all previous datasets, as shown in Fig. 1(b). This is especially helpful for training models for practical applications in which both high- and low-resolution human instances can be captured because the distance between the camera and the subject and the weather condition can both vary. We summarize the main advantages of MEBOW over previous HBOE datasets in Table 3.2.

Annotation tool. The annotation tool we used for labeling body orientation is illustrated in Fig. A1 of Appendix A. On the left side, one image from the dataset containing

Dataset	# subjects	# bins	Diversity	Occlusion
TUD*[6]	$5K$	8	✓	✓
3DPes[7]	$1K$	8	✗	✗
CASIA[41]	$19K$	6	✗	✗
MEBOW	$130K$	72	✓✓✓	✓✓✓

Table 1. Comparison of previous HBOE datasets with MEBOW. *Continuous body orientation labels of TUD are provided by [18].

human body instance(s) is displayed on the top. The associated cropped human instances is displayed at the bottom, from which the labeler could select which human instance to label by a mouse click. In the middle, the selected cropped human instance is displayed. On the right side, a slider is provided to adjust the orientation label in the range of $[0°, 360°)$ (default $0°$, step size $5°$), together with a clock-like circle and a red arrow visualizing the current labeled orientation. The labeler could first mouse-adjust the slider for coarse-grained orientation selection and then click either the *clock-wise++* or *counter clock-wise++* button (or using associated keyboard shortcuts) for fine-grained adjustments. The red arrow serves as a visual reference such that the labeler can compare it with the human body in the middle to ensure that the final orientation label is an accurate record of his/her comprehension. To maximize label consistency, on the bottom right corner, the labeler can refer to some example human body instances already labeled with the same orientation the labeler current selects.

Evaluation method. Given our high-precision 72-bin annotation, we propose to add Accuracy-$5°$, Accuracy-$15°$, and Accuracy-$30°$ as new evaluation metrics, where Accuracy-$X°$ is defined as the percentage of the samples that are predicted within $X°$ from the ground-truth orientation. As discussed in [18], mean absolute error (MAE) of the angular distance can be strongly influenced by a few large errors. However, Accuracy-$X°$ is less sensitive to the outliers, hence deserves more attentions as an evaluation criterion.

3.3. Baseline HBOE Model

Just as most previous work in HBOE, our baseline model assumes the human instances are already detected and the input is a cropped-out human instance. The cropping could be based on either the ground truth or the predicted bounding box. And for the ease of experiments, we used the ground-truth bounding boxes provided by the COCO dataset in all of our experiments. The overall network architecture of our baseline model is shown in Fig. 3(a), which can be trained end-to-end. The cropped images of subjects are first processed through a backbone network as the feature extractor. The extracted features are then concatenated and processed by a few more residual layers, with one fully connected layer and a softmax layer at the end. The output are 72 neurons, $\mathbf{p} = [p_0, p_2, ..., p_{71}]$ ($\sum_{i=0}^{71} p_i = 1.0$), rep-
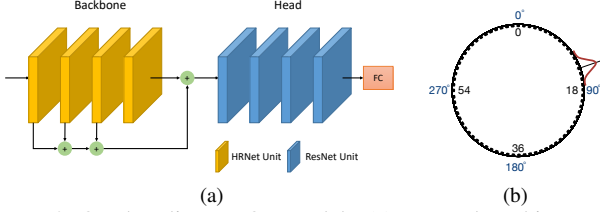
Figure 3. Our baseline HBOE model. (a) Network architecture. We adopt HRNet and ResNet units as the backbone network and the head network, respectively. Intermediate feature representations are combined to feed into the head network.(b) Illustration of 72 orientation bins (black ticks) and our orientation loss for regressing **p** to the "circular" Gaussian target probability function.

resenting the probability of every possible orientation bin being the best one to represent the body orientation of the input image. More specifically, p_i represents the probability of the body orientation θ to be within the i-th bin in Fig. 3(b), *i.e.*, within the range of $[i \cdot 5° - 2.5°, i \cdot 5° + 2.5°]$. As for the objective function of the model, our approach is different from previous approaches that either directly regress the orientation parameter θ (Approach 1 and 2 of [19]) or treat the orientation estimation as a pure classification problem (Approach 3 of [19], and [18]), where each bin is a different class. Instead, we take inspiration from the heat map regression idea, which has been extremely successful in key-point estimation [34, 46], and let the loss function for **p** be:

$$\mathcal{L} = \sum_{i=0}^{71} (p_i - \phi(i, \sigma))^2 , \qquad (1)$$

where $\phi(i, \sigma)$ is the "circular" Gaussian probability, as illustrated in Fig. 3(b) (red curve):

$$\phi(i, \sigma) = \frac{1}{\sqrt{(2\pi)}\sigma} e^{-\frac{1}{2\sigma^2}(min(|i-l_{gt}|, 72-|i-l_{gt}|))^2} , \quad (2)$$

and l_{gt} is the ground-truth orientation bin. Basically, we are regressing a Gaussian function centered at the ground truth orientation bin. And the intuition behind this is that the closer a orientation bin is to the ground-truth orientation bin label l_{gt}, the higher the probability the model should assign to it. We found this approach significantly eased the learning process of the neural network. And of note, we have attempted to use standard classification loss function, *e.g.* cross entropy loss between **p** and the ground truth represented by one hot vector, but the loss could not converge.

Choice of network architecture. We also considered ResNet-50 and ResNet-101 (initialized from the weights of the model trained for ImageNet classification task) to be the architecture of our network. We observe that HRNet+Head provides much better performance in experiments. This

could be explained by the fact that the HRNet and its pretrained model are also trained on COCO images and designed for a closer related task—2-D pose estimation.

3.4. Enhancing 3-D Pose Estimation

It is extremely difficult to obtain 3-D joint labels using existing technologies, hence models trained on indoor 3-D pose dataset generalize poorly to in-the-wild images, such as COCO images. There have been attempts [47, 48] to leverage 2-D pose datasets, such as MPII and COCO, as a second source of supervision to enhance both the performance and robustness of 3-D pose models. We believe the orientation labels in our COCO-based dataset can be complementary to the 2-D pose labels and provide additional supervision. To that end, we developed a triple-source weakly-supervised solution for 3-D pose estimation, the core of which is a *body orientation loss* for utilizing the orientation labels.

We choose [48] as the base to build our model. Following their notation, we denote $\mathbf{p} = [p_x, p_y, p_z]$ ($p_x \in [1, W], p_y \in [1, H], p_z \in [1, D]$) to be the coordinate of any location, and $\hat{\mathbf{H}}_k$ (of size $W \times H \times D$) to be the normalized heat map for joint k output by the backbone network. Then, the predicted location for joint k is:

$$\hat{\mathbf{J}}_k = \sum_{p_z=1}^{D} \sum_{p_y=1}^{H} \sum_{p_x=1}^{W} \mathbf{p} \cdot \hat{\mathbf{H}}_k(\mathbf{p}) . \qquad (3)$$

Next, $\mathcal{L}_2$ loss $L_{3D} = ||\hat{\mathbf{J}}_k - \mathbf{J}_k||^2$ can be used to supervise the network for images with 3-D pose labels. For images with 2-D pose labels, 1-D x heat vector and y heat vector is computed as:

$$\hat{\mathbf{J}}\hat{\mathbf{h}}_k^x = \sum_{p_x=1}^{W} \mathbf{p} \cdot \left(\sum_{p_z=1}^{D} \sum_{p_y=1}^{H} \hat{\mathbf{H}}_k(\mathbf{p}) \right) , \qquad (4)$$

$$\hat{\mathbf{J}}\hat{\mathbf{h}}_k^y = \sum_{p_y=1}^{H} \mathbf{p} \cdot \left(\sum_{p_z=1}^{D} \sum_{p_x=1}^{W} \hat{\mathbf{H}}_k(\mathbf{p}) \right) . \qquad (5)$$

And $\mathcal{L}_2$ loss $L_{2D} = ||\hat{\mathbf{J}}\hat{\mathbf{h}}_k^x - \mathbf{J}_k^x||^2 + ||\hat{\mathbf{J}}\hat{\mathbf{h}}_k^y - \mathbf{J}_k^y||^2$ can be used to supervise the network for images with 2-D pose labels.

Let's define the loss function for images with orientation labels. For the ease of notation, we use $\hat{\mathbf{J}}_{ls}, \hat{\mathbf{J}}_{rs}, \hat{\mathbf{J}}_{lh}$, and $\hat{\mathbf{J}}_{rh}$ to denote the predicted coordinates of *left shoulder*, *right shoulder*, *left hip*, and *right hip*, respectively, by Eq. 3. Then the estimated shoulder vector $\hat{\mathbf{S}}$ and torso vector $\hat{\mathbf{T}}$ can be represented by:

$$\hat{\mathbf{S}} = \hat{\mathbf{J}}_{rs} - \hat{\mathbf{J}}_{ls} , \qquad (6)$$

$$\hat{\mathbf{T}} = \frac{1}{2}(\hat{\mathbf{J}}_{lh} + \hat{\mathbf{J}}_{rh} - \hat{\mathbf{J}}_{lh} - \hat{\mathbf{J}}_{rh}) , \qquad (7)$$

following the definition in Sec. 2 and Fig. 2. And the chest facing direction can be computed by

$$\hat{\mathbf{C}} = \frac{\hat{\mathbf{T}} \times \hat{\mathbf{S}}}{||\hat{\mathbf{T}} \times \hat{\mathbf{S}}||_2} \ , \tag{8}$$

where $|| \cdot ||_2$ is the Euclidean norm. Since the (estimated) orientation angle $\hat{\theta}$ defined in Fig. 2 can be computed by projecting $\hat{\mathbf{C}}$ onto the y-z plane, we know the following equations hold:

$$\cos(\hat{\theta}) = \hat{\mathbf{C}}^z \ , \tag{9}$$

$$\sin(\hat{\theta}) = \hat{\mathbf{C}}^y \ . \tag{10}$$

And we define the *orientation loss* to be:

$$L_{\text{ori}} = ||\hat{\mathbf{C}}^z - \cos(\theta)||^2 + ||\hat{\mathbf{C}}^y - \sin(\theta)||^2 \ , \tag{11}$$

where θ is the ground truth orientation label. Finally, L_{2D}, L_{3D}, and L_{ori} can be used jointly with proper weighting between them such that the three sources of supervision, *i.e.*, 2-D pose labels, 3-D pose labels, and orientation labels, can all be used towards training a robust 3-D pose estimation model.

4. Experimental Results

The proposed MEBOW dataset has been tested in two sets of experiments for demonstrating its usefulness. In Sec. 4.1, we show how MEBOW can help advance HBOE by using the baseline model we proposed in Sec. 3.3. In Sec. 4.2, we show how MEBOW can help improve 3-D body pose estimation by using the triple-source weakly-supervised solution described in Sec. 3.4.

Implementation. All the codes used in the experiments were implemented with PyTorch [1]. For the HBOE experiments in Sec. 4.1, The ResNet backbone is based on the public codes [2], and is initialized from an ImageNet pre-trained model. The HRNet backbone is based on the public codes [3], and is initialized from a pretrained model for COCO 2-D pose estimation. The same input image preprocessing steps for the MEBOW and TUD datasets are applied, including normalizing the input images to 256×192, and flipping and scaling augmentation. We use Adam optimizer (learning rate = 1e−3) to train the network for 80 epochs. For the 3-D pose estimation experiments described in Sec. 4.2, our codes are based on public codes [4]. The network is initialized from an ImageNet pretrained model. Input images are normalized to 256×256. Rotation, flipping, and scaling are used to augment Human3.6M and MPII. To avoid the deformation of orientation, we do not carry out rotation augmentation for the images in MEBOW. The network is trained for 300 epochs. The Adam is the optimizer. The learning rate remains 1e−3.

4.1. Body Orientation Estimation

First, we validate the baseline model we proposed in Sec. 3.3. Specifically, we train it on the TUD dataset and compare its performance with other state-of-the-art models reported in the literature. The results are shown in Table 2. Our model significantly outperforms all of other models in terms of MAE, Accuracy-22.5°, and Accuracy-45°, which are standard metrics on the TUD dataset. This could be attributed to both our novel loss function for regressing the target "circular" Gaussian probability and the power of HRNet [46] and its pretrained model.

Method	MAE	Acc.-22.5°	Acc.-45°
AKRF-VW [18]	34.7	68.6	78
DCNN [19]	26.6	70.6	86.1
CPOEHK [53]	15.3	75.7	96.8
ours	8.4	95.1	99.7
Human [18]	0.93	90.7	99.3

Table 2. HBOE evaluation on the TUD dataset (with continuous orientation label). **Ours** was trained on the TUD training set and evaluated on its test set. We converted the continuous orientation label to 72-bin orientation label illustrated in Fig. 3.

To show the advantage of MEBOW over TUD in terms of diverse background and rich in-the-wild environment, we train our baseline model under four settings to compare the generalization capability of the same architecture (our proposed baseline model) trained on TUD and MEBOW. Our experimental results are shown in Table 3. It can be seen that the performance drop of our baseline model *trained on the TUD training set when it is evaluated on the MEBOW test set versus on the TUD test set* is much higher than that of the same model *trained on the MEBOW training set when it is evaluated on the TUD test set versus on the MEBOW test set*. This suggests that the improved diversity, and the inclusion of more challenging cases in MEBOW (compared with TUD) actually helps improve the robustness of models. We observe that Accuracy-45° for our model trained on MEBOW even improved slightly when evaluated on TUD versus on MEBOW. We also observe that the performance of our model, which is only trained on MEBOW (row 4 Table 3), can even exceed the previous state-of-the-art result on TUD (row 3 Table 2). Experiments of similar fashion and motivation have been conducted in Sec. 7 (Table. 1) of [26] to demonstrate the advantage of the COCO dataset.

Training	Testing	MAE	Acc.-22.5°	Acc.-45°
TUD	TUD	8.4	95.1	99.7
TUD	MEBOW	$32.2_{+23.8}$	$49.7_{-45.4}$	$77.5_{-22.2}$
MEBOW	MEBOW	8.4	93.9	98.2
MEBOW	TUD	$14.3_{+5.9}$	$77.3_{-16.6}$	$99.0_{+0.8}$

Table 3. Comparison of the generalization capability of the same model trained on TUD and on MEBOW.

As for the choice of the network architecture and the parameter σ, we conducted ablation experiments for both of them, with the results summarized in Table 4. HRNet+Head (initialized with pretrained weights for the COCO 2-D pose estimation task) gives significant better results than ResNet-50 or ResNet-101. And setting $\sigma = 4.0$ leads to the best performing model. Hence, we used the model with the HRNet+Head and $\sigma = 4.0$ for experiments associated with Table 2 and Table 3. Some qualitative prediction examples of this model are presented in Fig. 4.

Architecture	σ	MAE	Acc.-5°	Acc.-15°	Acc.-30°
ResNet-50	4.0	10.465	66.9	88.3	94.6
ResNet-101	4.0	10.331	67.8	88.2	94.7
	1.0	8.579	69.3	89.6	96.4
	2.0	8.529	**69.6**	**91.0**	96.6
	3.0	8.427	69.3	90.6	96.7
HRNet+Head	4.0	**8.393**	68.6	90.7	**96.9**
	6.0	8.556	68.2	90.9	96.7
	8.0	8.865	66.5	90.1	96.6

Table 4. Ablation study on the choice of network architecture and the effect of different σ in Eq. 2. Evaluation is done on MEBOW.

Figure 4. HBOE results generated by our baseline model (with HRNet as the backbone and $\sigma = 4.0$) on MEBOW (row 2, for respective images in row 1) and TUD dataset (row 5 for respective images in row 3). Row 4 are prediction results by [18] and they are directly cropped from the original paper. Red arrow: ground truth; Blue arrow: prediction.

4.2. Enhanced 3-D Body Pose Estimation

Data. we use the Human3.6M dataset (3-D pose), the MPII dataset (2-D pose), the COCO dataset (2-D pose), and our MEBOW orientation labels. We train our triple-source weakly-supervised model proposed in Sec. 3.4 and two dual-source weakly-supervised baseline models for comparison. Both of the baseline models are trained using a re-implementation of [48], which uses a combination of $L_{2D} + L_{3D}$ (defined in Sec. 3.4). The difference is that the Baseline-1 only uses Human3.6M dataset (3-D pose) and the MPII dataset (2-D pose), while the Baseline-2 uses COCO dataset (2-D pose) on top of the first baseline. Our method leverages the orientation labels from our MEBOW dataset on top of the second baseline and uses a combination of $L_{2D} + L_{3D} + L_{ori}$. Following the practice of [48], within a batch during the stochastic training, we sampled the same number of images from Human3.6, MPII, and COCO datasets.

We evaluated and compared our model and the two baselines in multiple ways, both quantitatively and qualitatively. *First*, we followed the Protocol II in [48] and used the mean per joint position error (MPJPE) as the metric to evaluate them on the test set of the Human3.6M dataset. The evaluation results are shown in Table 5, along with the evaluation results for other competitive models copied from their papers. We have tried our best to train Baseline-1 but still cannot obtain a model with a performance as good as that reported in [48]. This, however, does not hinder us from making a fair comparison between Baseline-1, Baseline-2, and our model. From Table 5, we can see that by adding MEBOW as the third (weak) supervision source and using our proposed *orientation loss* L_{ori}, we can achieve significantly better average MPJPE than both Baseline-1 and Baseline-2. If we break down MPJPE metric into different motion categories, our approach also achieves the best MPJPE metric in most (12 out of 16) motion categories. We also did breakdown analysis of the MPJPE metric in terms of different joints and X-, Y-, Z- part of the joint coordinates in Table 6. For nearly all joints, our method achieves significant better results. And our method is positive on improving Y- and Z- part of the joint coordinate but neutral for improving X- part of the joint coordinate. This is not surprising since our *orientation loss* only considers the Y- and Z- part of **C** after the projection on to the y-z plane in Fig. 2. Some qualitative examples of 3-D pose estimation by our model, along with the ground truth and the predictions by the two baseline models are displayed in Fig. 5. *Second*, we conduct evaluation of the 3-D pose prediction on the COCO test set. Since the ground-truth 3-D pose is unknown for the COCO dataset, we took a step back and conducted the quantitative evaluation by comparing the orientation computed from the predicted 3-D pose against the ground-truth orientation label provided by our MEBOW dataset. As shown in Table 7, our model significantly outperforms both Baseline-1 and Baseline-2, which suggests our model for 3-D pose estimation generalizes much better to in-the-wild images. Fig. 6 shows a few qualitative results of 3-D pose prediction on the COCO test set.

5. Conclusions

We introduced a new COCO-based large-scale, high-precision dataset for human body orientation estimation in

Method	Dir.	Dis.	Eat.	Gre.	Phon.	Pose	Pur.	Sit	SitD.	Smo.	Phot.	Wait	Walk	WalkD.	WalkP.	Average
Chen *et al.* [11]	89.9	97.6	90.0	107.9	107.3	139.2	93.6	136.1	133.1	240.1	106.7	106.2	87.0	114.1	90.6	114.2
Tome *et al.* [50]	65.0	73.5	76.8	86.4	86.3	110.7	68.9	74.8	110.2	172.9	85.0	85.8	86.3	71.4	73.1	88.4
Zhou *et al.* [56]	87.4	109.3	187.1	103.2	116.2	143.3	106.9	99.8	124.5	199.2	107.4	118.1	114.2	79.4	97.7	79.9
Metha *et al.* [30]	59.7	69.7	60.6	68.8	76.4	85.4	59.1	75.0	96.2	122.9	70.8	68.5	54.4	82.0	59.8	74.1
Pavlakos *et al.* [37]	58.6	64.6	63.7	62.4	66.9	70.8	57.7	62.5	76.8	103.5	65.7	61.6	67.6	56.4	59.5	66.9
Moreno *et al.* [32]	69.5	80.2	78.2	87.0	100.8	102.7	76.0	69.7	104.7	113.9	89.7	98.5	82.4	79.2	77.2	87.3
Sun *et al.* [47]	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
Sharma *et al.* [44]	48.6	54.5	54.2	55.7	62.6	72.0	50.5	54.3	70.0	78.3	58.1	55.4	61.4	45.2	49.7	58.0
Moon *et al.* [31]	50.5	55.7	50.1	51.7	53.9	46.8	50.0	61.9	68.0	52.5	55.9	49.9	41.8	56.1	46.9	53.3
Sun *et al.* [48]	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
Baseline-1*	44.4	47.4	49.0	67.7	50.0	41.8	45.6	59.9	92.9	48.8	57.1	65.4	38.7	50.5	42.2	53.4
Baseline-2**	46.1	47.8	49.1	66.3	48.0	43.5	46.7	59.3	85.0	47.0	54.0	61.9	38.6	50.1	49.7	52.4 $_{-1.0}$
ours	44.6	47.1	46.0	60.5	47.7	41.8	46.0	57.8	82.3	47.2	56.0	56.7	38.0	49.5	41.8	50.9 $_{-2.5}$

Table 5. 3-D human pose estimation evaluation on the Human3.6M dataset using mean per joint position error (MPJPE). *Our baseline is a re-implementation of Sun *et al.* [48], trained on Human3.6M + MPII, as in the original paper. **Our baseline 2 is a re-implementation of Sun *et al.* [48], trained on Human3.6M + MPII + COCO (2-D Pose). The best and second best are marked with color.

Method	Hip$^+$	Knee$^+$	Ankle$^+$	Torso	Neck	Head	Nose	Shoulder$^+$	Elbow$^+$	Wrist$^+$	X	Y	Z (Depth)
Baseline-1*	24.6	49.0	73.8	40.6	51.9	55.6	56.9	52.5	66.8	84.8	14.6	19.4	39.8
Baseline-2**	23.5$_{-1.1}$	49.7$_{+0.7}$	72.6$_{-1.2}$	36.8$_{-3.8}$	50.4$_{-1.5}$	53.0$_{-2.6}$	49.6$_{-7.3}$	51.0$_{-1.5}$	66.0$_{-0.8}$	87.6$_{+2.8}$	14.3$_{-0.3}$	18.2$_{-1.2}$	39.8$_{+0.0}$
ours	21.6$_{-3.0}$	45.7$_{-3.3}$	68.9$_{-4.9}$	35.2$_{-5.4}$	47.9$_{-4.0}$	51.1$_{-4.5}$	52.3$_{-4.6}$	49.6$_{-2.9}$	65.9$_{-0.9}$	87.6$_{+2.8}$	14.7$_{+0.1}$	17.1$_{-2.3}$	39.0$_{-0.8}$

Table 6. 3-D human pose estimation **per joint** evaluation on the Human3.6M dataset using mean per joint position error (MPJPE). $^+$The error is the average of the left joint and the right joint.
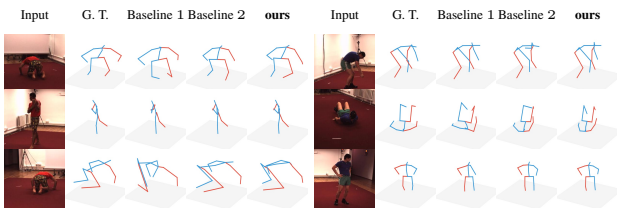

Figure 5. Example 3-D pose estimation results on the Human3.6M dataset. (G.T. is the abbreviation for Ground Truth.) More example results can be viewed in Appendix E.
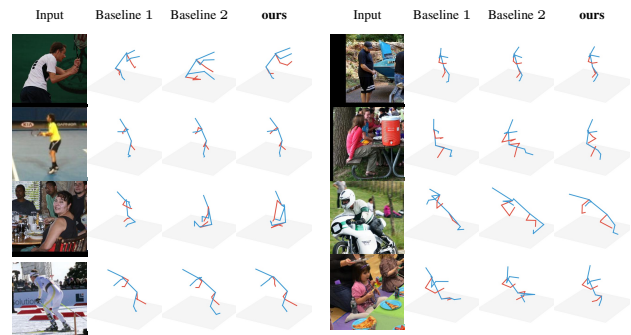

Figure 6. Example 3-D pose estimation results on the COCO dataset. More example results can be viewed in Appendix F.

the wild. Through extensive experiments, we demonstrated that our dataset could be very useful for both body orientation estimation and 3-D pose estimation. In the meanwhile, we presented a simple, yet effective model for human body orientation estimation, which can serve as a baseline for future HBOE model development using our dataset. And we proposed a new *orientation loss* for utilizing body orientation label as the third supervision source. In the future, it would be interesting to explore how our dataset can be used for other vision tasks, *e.g.*, person re-identification (ReID) and bodily expressed emotion recognition [29].

Acknowledgments

Method	MAE	Acc.-5°	Acc.-15°	Acc.-30°
Baseline*	26.239	34.7	63.7	77.7
Baseline 2**	13.888	31.9	74.5	86.8
ours	11.023	44.8	83.4	94.2

Table 7. 3-D human pose estimation evaluation on the test portion of MEBOW.

References

[1] https://pytorch.org. 6

[2] https://github.com/pytorch/vision/blob/master/\protect\penalty-\@Mtorchvision/models/resnet.py. 6

[3] https://github.com/leoxiaobin/deep-high-resolution-net.pytorch. 6

[4] https://github.com/JimmySuen/integral-human-pose. 6

[5] 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*, June 2014. 3

[6] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Monocular 3D pose estimation and tracking by detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 623–630. IEEE, 2010. 1, 2, 3, 4

[7] Davide Baltieri, Roberto Vezzani, and Rita Cucchiara. 3DPeS: 3D people dataset for surveillance and forensics. In *Proceedings of the Joint ACM Workshop on Human Gesture and Behavior Understanding*, pages 59–64. ACM, 2011. 3, 4

[8] Davide Baltieri, Roberto Vezzani, and Rita Cucchiara. People orientation recognition by mixtures of wrapped distributions on random trees. In *Proceedings of the European Conference on Computer Vision*, pages 270–283. Springer, 2012. 2, 3

[9] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Proceedings of the European Conference on Computer Vision*, pages 561–578. Springer, 2016. 3

[10] Cheng Chen, Alexandre Heili, and Jean-Marc Odobez. Combined estimation of location and body pose in surveillance video. In *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 5–10. IEEE, 2011. 2, 3

[11] 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*, pages 7035–7043, 2017. 8, 12

[12] Jinyoung Choi, Beom-Jin Lee, and Byoung-Tak Zhang. Human body orientation estimation using convolutional neural network. *arXiv preprint arXiv:1609.01984*, 2016. 3

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 2

[14] Tarak Gandhi and Mohan Manubhai Trivedi. Image based estimation of pedestrian orientation for improving path prediction. In *Proceedings of the IEEE Intelligent Vehicles Symposium*, pages 506–511. IEEE, 2008. 2, 3

[15] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 2

[16] Amir Ghodrati, Marco Pedersoli, and Tinne Tuytelaars. Is 2D information enough for viewpoint estimation? In *Proceedings of the British Machine Vision Conference*, volume 101, page 102, 2014. 3

[17] Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. Estimating human shape and pose from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1381–1388. IEEE, 2009. 3

[18] Kota Hara and Rama Chellappa. Growing regression tree forests by classification for continuous object pose estimation. *International Journal of Computer Vision*, 122(2):292–312, 2017. 2, 3, 4, 5, 6, 7

[19] Kota Hara, Raviteja Vemulapalli, and Rama Chellappa. Designing deep convolutional neural networks for continuous object orientation estimation. *arXiv preprint arXiv:1702.01499*, 2017. 2, 3, 5, 6

[20] 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*, 36(7):1325–1339, 2013. 3

[21] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 3

[22] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 3

[23] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2252–2261, 2019. 3

[24] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3D and 2D human representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6050–6059, 2017. 3

[25] Jiahao Lin and Gim Hee Lee. Trajectory space factorization for deep video-based 3D human pose estimation. 2019. 3

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014. 2, 3, 4, 6

[27] Peiye Liu, Wu Liu, and Huadong Ma. Weighted sequence loss based spatial-temporal deep learning framework for human body orientation estimation. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 97–102. IEEE, 2017. 2, 3

[28] Wu Liu, Yongdong Zhang, Sheng Tang, Jinhui Tang, Richang Hong, and Jintao Li. Accurate estimation of human body orientation from rgb-d sensors. *IEEE Transactions on Cybernetics*, 43(5):1442–1452, 2013. 3

[29] Yu Luo, Jianbo Ye, Reginald B Adams, Jia Li, Michelle G Newman, and James Z Wang. Arbee: Towards automated recognition of bodily expression of emotion in the wild. *International Journal of Computer Vision*, 128(1):1–25, 2019. 8

[30] 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. *arXiv preprint arXiv:1611.09813*, 1(3):5, 2016. 3, 8

[31] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3D multi-person pose estimation from a single rgb image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10133–10142, 2019. 8, 12

[32] 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*, pages 2823–2832, 2017. 8, 12

[33] Chikahito Nakajima, Massimiliano Pontil, Bernd Heisele, and Tomaso Poggio. Full-body person recognition system. *Pattern Recognition*, 36(9):1997–2006, 2003. 2, 3

[34] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Proceedings of the European Conference on Computer Vision*, pages 483–499. Springer, 2016. 5

[35] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 484–494. IEEE, 2018. 3

[36] 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*, pages 7307–7316, 2018. 3

[37] 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*, pages 7025–7034, 2017. 8

[38] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 459–468, 2018. 3

[39] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7753–7762, 2019. 3

[40] Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. Cross view fusion for 3D human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4342–4351, 2019. 3

[41] Rahul Raman, Pankaj Kumar Sa, Banshidhar Majhi, and Sambit Bakshi. Direction estimation for pedestrian monitoring system in smart cities: An hmm based approach. *IEEE Access*, 4:5788–5808, 2016. 3, 4

[42] Mudassar Raza, Zonghai Chen, Saeed-Ur Rehman, Peng Wang, and Peng Bao. Appearance based pedestrians head pose and body orientation estimation using deep learning. *Neurocomputing*, 272:647–659, 2018. 3

[43] Grégory Rogez and Cordelia Schmid. Mocap-guided data augmentation for 3D pose estimation in the wild. In *Advances in Neural Information Processing Systems*, pages 3108–3116, 2016. 3

[44] Saurabh Sharma, Pavan Teja Varigonda, Prashast Bindal, Abhishek Sharma, and Arjun Jain. Monocular 3D human pose estimation by generation and ordinal ranking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2325–2334, 2019. 8, 12

[45] Hiroaki Shimizu and Tomaso Poggio. Direction estimation of pedestrian from multiple still images. In *Proceedings of the IEEE Intelligent Vehicles Symposium*, pages 596–600. IEEE, 2004. 2, 3

[46] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 5, 6

[47] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2602–2611, 2017. 5, 8, 12

[48] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision*, pages 529–545, 2018. 3, 5, 7, 8, 12

[49] Bugra Tekin, Pablo Márquez-Neila, Mathieu Salzmann, and Pascal Fua. Learning to fuse 2D and 3D image cues for monocular body pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3941–3950, 2017. 3

[50] 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*, pages 2500–2509, 2017. 8

[51] John Towns, Timothy Cockerill, Maytal Dahan, Ian Foster, Kelly Gaither, Andrew Grimshaw, Victor Hazlewood, Scott Lathrop, Dave Lifka, Gregory D Peterson, et al. XSEDE: Accelerating scientific discovery. *Computing in Science & Engineering*, 16(5):62–74, 2014. 8

[52] 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*, pages 4948–4956, 2016. 3

[53] Dameng Yu, Hui Xiong, Qing Xu, Jianqiang Wang, and Keqiang Li. Continuous pedestrian orientation estimation using human keypoints. In *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE, 2019. 2, 3, 6

[54] Guangzhe Zhao, Mrutani Takafumi, Kajita Shoji, and Mase Kenji. Video based estimation of pedestrian walking direction for pedestrian protection system. *Journal of Electronics (China)*, 29(1-2):72–81, 2012. 2, 3

[55] 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*, pages 398–407, 2017. 3

[56] 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*, 41(4):901–914, 2018. 8, 12

CVPR Supplementary Material

A. Labeling Tool

The interface of our human body orientation labeling tool is illustrated in Fig. A1.
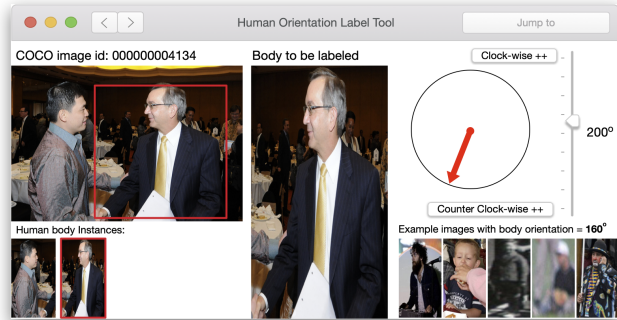

Figure A1. User interface of the labeling tool.

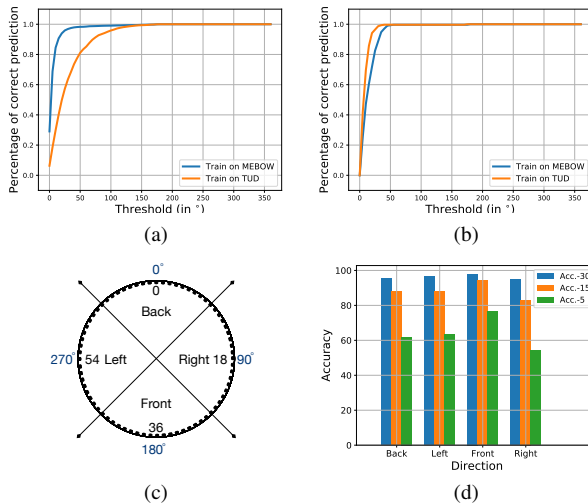## B. More Details on the HBOE Experiments

Figure A2. Breakdown analysis of the performance of our HBOE baseline model.

Breakdown analysis of the errors. *First*, we show the cumulative percentage of correct HBOE prediction with respect to the threshold of a correct prediction in Fig. A2 (a) and (b). Specifically, we compare the performance of our baseline model trained on the MEBOW dataset and that trained on the TUD dataset, respectively, using 1) the test set of the MEBOW dataset (Fig. A2 (a)) and 2) the test set of the TUD dataset (Fig. A2 (b)). Based on the same set of experiments in Table 3, these two sub-figures present a more detailed comparison, and they also support that the model trained on our MEBOW dataset has much better generalizability than it trained on TUD dataset. *Second*, we show how our baseline HBOE model performs when the camera point of view is towards the *Front*, *Back*, *Left*, and

Right of the person in Fig. A2 (d). The association of the ground-truth orientations with the *Front*, *Back*, *Left*, and *Right* breakdown categories are shown in Fig. A2 (c). It is not surprising that our model performs best when the camera point of view is towards the *Front* of the person because a larger portion of MEBOW dataset falls into this category, as shown in Fig. 1 (a) in the main paper.

C. Additional 3-D Human Pose Estimation Evaluation on the Human3.6M Dataset

We also conducted 3-D human pose estimation experiments with Protocol I in [48]. The evaluation results are shown in Tabel A1.

Method	PA MPJPE
Chen *et al.* [11]	82.7
Moreno *et al.* [32]	76.5
Zhou *et al.* [56]	55.3
Sun *et al.* [47]	48.3
Sharma *et al.* [44]	40.9
Sun *et al.* [48]	40.6
Moon *et al.* [31]	34.0
Baseline*	34.7
Baseline 2**	34.3
ours	33.1

Table A1. 3-D human pose estimation evaluation on the Human3.6M dataset using Protocol I. *Our baseline is a re-implementation of Sun *et al.* [48], trained on Human3.6M + MPII, as in the original paper. **Our baseline 2 is a re-implementation of Sun *et al.* [48], trained on Human3.6M + MPII + COCO (2-D Pose).

D. More Qualitative Human Body Orientation Estimation Results

More qualitative human body orientation estimation examples are shown in Fig. A3 to supplement Fig. 4 in the main paper.

E. More Qualitative 3-D Pose Estimation Results on the Human3.6M Dataset

More example 3-D pose estimation results on the test set of the Human3.6M dataset are included in Fig. A4.

F. More Qualitative 3-D Pose Estimation Results on the COCO Dataset

More example 3-D pose estimation results on the test set of the COCO dataset are shown in Fig. A5.
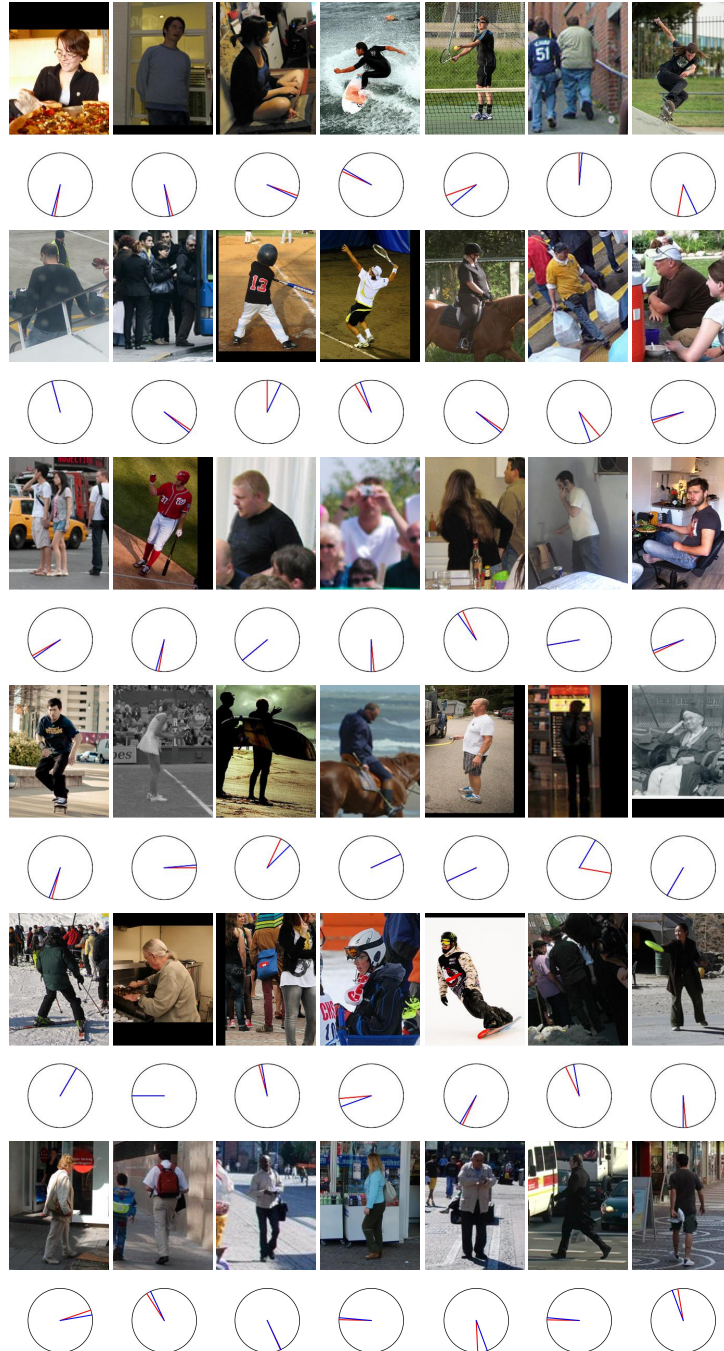
Figure A3. HBOE results generated by our baseline model (with HRNet as the backbone and $\sigma = 4.0$) on MEBOW (row 1 to row 5) and TUD dataset (row 6). Red arrow: ground truth; Blue arrow: prediction.
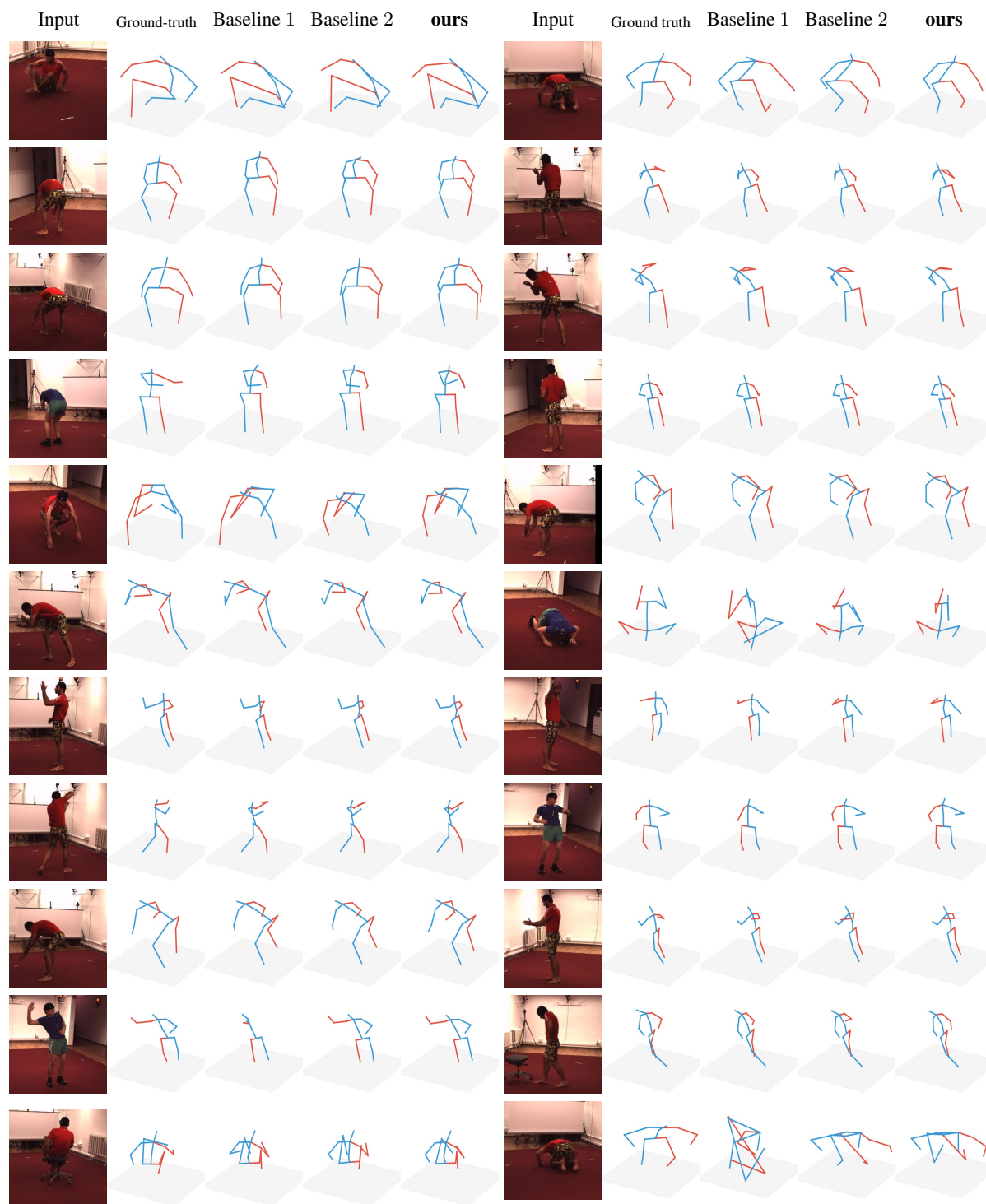
Figure A4. More example 3-D pose estimation results on the Human3.6M dataset.

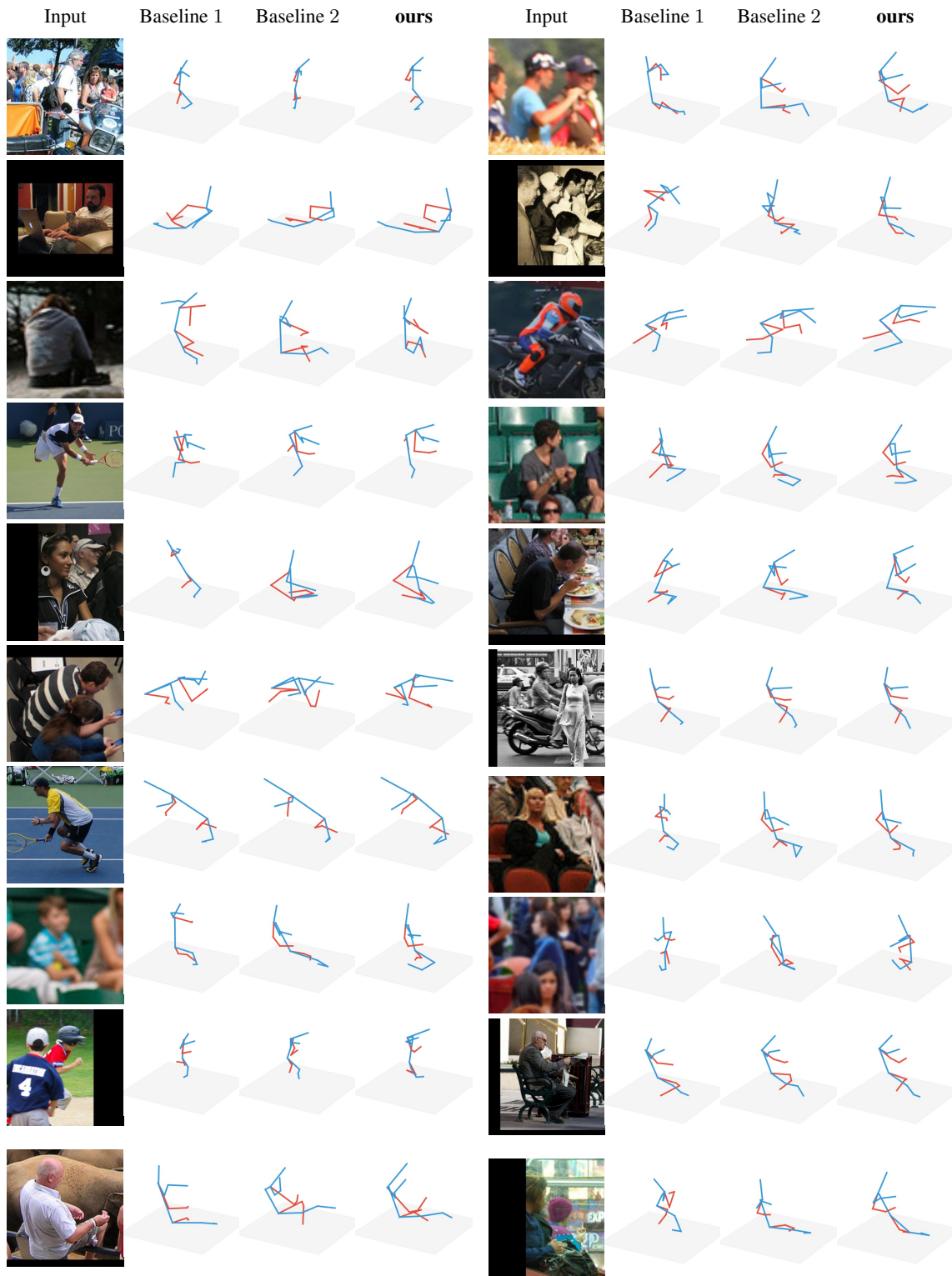| Input	Baseline 1	Baseline 2	**ours**	Input	Baseline 1	Baseline 2	**ours**

Figure A5. More example 3-D pose estimation results on the COCO dataset.