MEBOW: Monocular Estimation of Body Orientation In the Wild

Chenyan Wu¹ ²*, Yukun Chen¹ Jiajia Luo² Che-Chun Su² Anuja Dawane²
Bikramjot Hanzra² Zhuo Deng² Bilan Liu² James Z. Wang¹ Cheng-hao Kuo²
¹ The Pennsylvania State University, 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, occluded and 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 value 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/](https://chenyanwu.github.io/MEBOW/).

1. Introduction

Human body orientation estimation (HBOE) aims at estimating 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, occluded and indistinguishable body parts. Second, HBOE is a more direct task towards understanding human behaviors. Third, HBOE is a simpler task compared with 3-D pose estimation and can be directly solved using pose estimation models. 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/](https://chenyanwu.github.io/MEBOW/).

*This work was mostly done when Chenyan Wu was an intern at Amazon Lab126.
able body parts, all of which are prevalent in in-the-wild images. Second, under certain scenarios, the orientation of the body is already good enough 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 to each other in addressing real-world challenges involving understanding human behaviors.

HBOE has been studied in recent years [18, 19, 51, 6, 44, 4, 32, 10, 52, 3, 21]. A primary bottleneck, however, is a lack of a large-scale, high-precision, diverse-background dataset. Previously, the TUD dataset [6] is 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 8 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), 3D 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 computer vision research, such as ImageNet [13] for image classification, KITTI [5] 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. 5.2. 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 [51] 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 could generalize 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 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, 51], 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 8 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 [40] (19, 139 frames of videos capturing 20 subjects) have been used in [51] and [41] [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 covered 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 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] [44] [14] [32] [10] [52] [8], e.g., SVM, have been favored for HBOE. Deep learning-based methods [41] [27] also treat HBOE as a classification problem. For example, the method in [41] uses a 14-layer classification network to predict which bin out of the 8 different bins represents the orientation given the input; the method in [27] uses a 4-layer neural network as the classification network. We can see that those 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 used for collecting training images. Given the continuous orientation label provided by [18] for the TUD dataset, some new works [18] [19] [51] attempt to address more fine-grained body orientation prediction. Most notably, Yu et al. [51] utilizes the key-points detection prediction by another 2-D pose model as an additional cue for the 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 2D images could outperform state-of-the-art methods based on 3D information (e.g., 3D CAD models or 3D 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 [50] [29] [42] [53] [48] [47] have been made to address this by using additional source of supervision, mainly 2-D pose dataset (e.g., MPII [51]). 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 includes: 1) enforcing extra prior knowledge such as a parameterized 3-D human mesh model [17] [24] [9] [23] [22] [34] [37], the ordinal depth [35], and temporal information (such as adjacent frame consistency) [25] [38]; 2) Leveraging images simultaneously captured from different views [39] [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.](image)

Previous datasets including TUD all assume that the human body orientation is self-explanatory from the image, which is fine 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 \( \theta \in [0^\circ, 360^\circ) \) to be 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 \( C = T \times S \), 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 left- and right-shoulder joints to the midpoint of 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 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. We can see that our dataset covers all possible body orientation, with a Gaussian like peak around 180°, which is natural because images 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 5.2.

**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 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 \([0^\circ, 360^\circ)\) (default \(0^\circ\), step size \(5^\circ\)), 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 \(\text{clock-wise++}\) or \(\text{counter clock-wise++}\) button (or using associated keyboard shortcuts) for fine-grained adjustment. 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-\(X^\circ\) is defined as the percentage of the samples that are predicted within \(X^\circ\) 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-\(X^\circ\) is less sensitive to the outliers and deserves more attentions as an evaluation criterion.

3.3. Baseline HBOE Model

Just as most previous HBOE works, 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 con-

<table>
<thead>
<tr>
<th>Dataset</th>
<th># subjects</th>
<th># bins</th>
<th>Diversity</th>
<th>Occlusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>TUD</td>
<td>0</td>
<td>5K</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>3DPes</td>
<td>1K</td>
<td>8</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>CASIA</td>
<td>19K</td>
<td>6</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>MEBOW</td>
<td>130K</td>
<td>72</td>
<td>✓ ✓ ✓</td>
<td>✓ ✓ ✓</td>
</tr>
</tbody>
</table>

Table 1. Comparison of previous HBOE datasets with MEBOW. *Continuous body orientation labels of TUD are provided by [18].
function for successful in key-point estimation [33, 45], and let the loss the heatmap regression idea, which has been extremely suc-

As for the objective function of the model, our approach i.e. being the best one to represent the body orientation of the presence weakly-supervised solution for 3-D pose estimation, the core of which is a body orientation loss for utilizing the orientation labels.

We choose [47] as the base to build our model. Following their notation, we denote \( p = [p_x, p_y, p_z] \) \( (p_z \in [1, W], p_y \in [1, H], p_z \in [1, D]) \) to be the coordinate of any location, and \( \hat{H}_k \) (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:

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

Then, \( \mathcal{L}_{3D} = ||\hat{J}_k - J_k||^2 \) 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:

\[
\hat{J}_{hx}^x = \sum_{p_z=1}^{W} \sum_{p_y=1}^{H} \sum_{p_z=1}^{D} p \cdot \hat{H}_k(p) ,
\] (4)

\[
\hat{J}_{hx}^y = \sum_{p_z=1}^{H} \sum_{p_y=1}^{W} \sum_{p_z=1}^{D} p \cdot \hat{H}_k(p) .
\] (5)

And \( \mathcal{L}_2 \) loss \( \mathcal{L}_{2D} = ||\hat{J}_{hx}^x - J_k^x||^2 + ||\hat{J}_{hx}^y - J_k^y||^2 \) 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 \( \hat{J}_{lx}, \hat{J}_{rs}, \hat{J}_{lh}, \) and \( \hat{J}_{rh} \) to denote the predicted coordinates of left shoulder, right shoulder, left hip, and right hip, respectively, by Eq. 3.

Then the estimated shoulder vector \( \hat{S} \) and torso vector \( \hat{T} \) can

**3.4. Enhancing 3D 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 [46, 47] 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.
be represented by:
\[
\hat{\boldsymbol{S}} = \hat{\boldsymbol{J}}_{rs} - \hat{\boldsymbol{J}}_{ls},
\]
\[
\hat{\mathbf{T}} = \frac{1}{2} (\hat{\boldsymbol{J}}_{lh} + \hat{\boldsymbol{J}}_{rh} - \hat{\boldsymbol{J}}_{lh} - \hat{\boldsymbol{J}}_{rh}),
\]
following the definition in Sec. 2 and Fig. 2. And the chest facing direction can be computed by
\[
\hat{\mathbf{C}} = \frac{\hat{\mathbf{T}} \times \hat{\mathbf{S}}}{||\hat{\mathbf{T}} \times \hat{\mathbf{S}}||_2},
\]
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:
\[
\cos(\hat{\theta}) = \hat{\mathbf{C}}^z, \\
\sin(\hat{\theta}) = \hat{\mathbf{C}}^y.
\]
And we define the orientation loss to be:
\[
L_{ori} = ||\hat{\mathbf{C}}^z - \cos(\theta)||^2 + ||\hat{\mathbf{C}}^y - \sin(\theta)||^2,
\]
where \( \theta \) 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, \text{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 showed how MEBOW can help advance HBOE by using the baseline model we proposed in Sec. 3.3. Specifically, we trained 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 [45] and its pretrained model.

<table>
<thead>
<tr>
<th>Method</th>
<th>MAE</th>
<th>Acc.-22.5°</th>
<th>Acc.-45°</th>
</tr>
</thead>
<tbody>
<tr>
<td>AKRF-VW</td>
<td>34.7</td>
<td>68.6</td>
<td>78</td>
</tr>
<tr>
<td>DCNN</td>
<td>26.6</td>
<td>70.6</td>
<td>86.1</td>
</tr>
<tr>
<td>CPOEHK</td>
<td>15.3</td>
<td>75.7</td>
<td>96.8</td>
</tr>
<tr>
<td>ours</td>
<td>8.4</td>
<td>95.1</td>
<td>99.7</td>
</tr>
<tr>
<td>Human</td>
<td>0.93</td>
<td>90.7</td>
<td>99.3</td>
</tr>
</tbody>
</table>

Table 2. HBOE evaluation on the TUD dataset (with continuous orientation label). \textbf{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 trained 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 displayed in Table 3. We can see that the performance drop of our baseline model \textit{trained on the TUD training set when it is evaluated on the MEBOW test set} versus \textit{on the TUD test set} is much higher than that of the same model \textit{trained on the MEBOW training set when it is evaluated on the TUD test set} versus \textit{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.

As for the choice of the network architecture and the parameter \( \sigma \), we did ablation experiments for both of them, with the results summarized in Table 4. We found HRNet + Head (initialized with pretrained weights for the
Table 3. Comparison of the generalization capability of the same model trained on TUD and on MEBOW.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>MAE</th>
<th>Acc.-22.5°</th>
<th>Acc.-45°</th>
</tr>
</thead>
<tbody>
<tr>
<td>TUD</td>
<td>8.4</td>
<td>95.1</td>
<td>99.7</td>
</tr>
<tr>
<td>TUD</td>
<td>32.2</td>
<td>23.8</td>
<td>49.7</td>
</tr>
<tr>
<td>MEBOW</td>
<td>8.4</td>
<td>93.9</td>
<td>98.2</td>
</tr>
<tr>
<td>MEBOW</td>
<td>14.3</td>
<td>5.9</td>
<td>77.3</td>
</tr>
</tbody>
</table>

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

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

COCO 2-D pose estimation task) gave significant better results than ResNet-50 or ResNet-101. And setting σ = 4.0 leads to the best performing model. Hence, we used the model with the HRNet+Head and σ = 4.0 for experiments associated with Table 2 and Table 3. Some qualitative prediction examples of this model are displayed in Fig. 4.

Figure 4. HBOE results generated by our baseline model (with HRNet as the backbone and σ = 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 trained 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 [47], 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 [47], within a batch during the stochastic training, we sampled the same number of images from Human3.6, MPII and COCO dataset.

We evaluated and compared our model and the two baselines in multiple ways, both quantitatively and qualitatively. First, we followed the Protocol II in [47] 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 [47]. 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 breakdown 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 project 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. From the results in Table 7, we can see our model significantly outperform 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 the 3-D pose prediction on the COCO test set.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chen et al.</td>
<td>11</td>
<td>89.9</td>
<td>97.6</td>
<td>90.0</td>
<td>107.9</td>
<td>107.3</td>
<td>139.2</td>
<td>93.6</td>
<td>136.1</td>
<td>133.1</td>
<td>240.1</td>
<td>106.7</td>
<td>106.2</td>
<td>87.0</td>
<td>114.1</td>
<td>90.6</td>
</tr>
<tr>
<td>Tome et al.</td>
<td>49</td>
<td>65.0</td>
<td>73.5</td>
<td>76.8</td>
<td>86.4</td>
<td>86.3</td>
<td>110.7</td>
<td>68.9</td>
<td>74.8</td>
<td>110.2</td>
<td>172.9</td>
<td>85.0</td>
<td>85.8</td>
<td>86.3</td>
<td>71.4</td>
<td>73.1</td>
</tr>
<tr>
<td>Zhou et al.</td>
<td>54</td>
<td>87.4</td>
<td>109.3</td>
<td>187.1</td>
<td>103.2</td>
<td>116.2</td>
<td>143.3</td>
<td>106.9</td>
<td>99.8</td>
<td>124.5</td>
<td>199.2</td>
<td>107.4</td>
<td>118.1</td>
<td>114.2</td>
<td>79.4</td>
<td>97.7</td>
</tr>
<tr>
<td>Metha et al.</td>
<td>29</td>
<td>59.7</td>
<td>69.7</td>
<td>60.6</td>
<td>68.8</td>
<td>76.4</td>
<td>85.4</td>
<td>59.1</td>
<td>75.0</td>
<td>96.2</td>
<td>122.9</td>
<td>70.8</td>
<td>68.5</td>
<td>54.4</td>
<td>82.0</td>
<td>59.8</td>
</tr>
<tr>
<td>Pavlakos et al.</td>
<td>36</td>
<td>58.6</td>
<td>64.6</td>
<td>63.7</td>
<td>62.4</td>
<td>66.9</td>
<td>70.8</td>
<td>57.7</td>
<td>62.5</td>
<td>76.8</td>
<td>103.5</td>
<td>65.7</td>
<td>61.6</td>
<td>67.6</td>
<td>56.4</td>
<td>59.5</td>
</tr>
<tr>
<td>Moreno et al.</td>
<td>31</td>
<td>69.5</td>
<td>80.2</td>
<td>78.2</td>
<td>87.0</td>
<td>100.8</td>
<td>102.7</td>
<td>76.0</td>
<td>69.7</td>
<td>104.7</td>
<td>113.9</td>
<td>89.7</td>
<td>98.5</td>
<td>82.9</td>
<td>79.2</td>
<td>77.2</td>
</tr>
<tr>
<td>Sun et al.</td>
<td>46</td>
<td>52.8</td>
<td>54.8</td>
<td>54.2</td>
<td>54.3</td>
<td>61.8</td>
<td>53.1</td>
<td>53.6</td>
<td>71.7</td>
<td>86.7</td>
<td>61.5</td>
<td>67.2</td>
<td>53.4</td>
<td>47.1</td>
<td>61.6</td>
<td>53.4</td>
</tr>
<tr>
<td>Sharma et al.</td>
<td>43</td>
<td>48.6</td>
<td>54.5</td>
<td>54.2</td>
<td>55.7</td>
<td>62.6</td>
<td>72.0</td>
<td>50.5</td>
<td>54.3</td>
<td>70.0</td>
<td>78.3</td>
<td>58.1</td>
<td>55.4</td>
<td>61.4</td>
<td>45.2</td>
<td>49.7</td>
</tr>
<tr>
<td>Moon et al.</td>
<td>30</td>
<td>50.5</td>
<td>55.7</td>
<td>50.1</td>
<td>51.7</td>
<td>53.9</td>
<td>46.8</td>
<td>50.0</td>
<td>61.9</td>
<td>68.0</td>
<td>52.5</td>
<td>55.9</td>
<td>49.9</td>
<td>41.8</td>
<td>56.1</td>
<td>46.9</td>
</tr>
<tr>
<td>Sun et al.</td>
<td>47</td>
<td>47.5</td>
<td>47.7</td>
<td>49.5</td>
<td>50.2</td>
<td>51.4</td>
<td>43.8</td>
<td>46.4</td>
<td>58.9</td>
<td>65.7</td>
<td>49.4</td>
<td>55.8</td>
<td>47.8</td>
<td>38.9</td>
<td>49.0</td>
<td>43.8</td>
</tr>
</tbody>
</table>

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 |
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 |

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. [47], trained on Human3.6M + MPII, as in the original paper. **Our baseline 2 is a re-implementation of Sun et al. [47], trained on Human3.6M + MPII + COCO (2-D Pose). The best and second best are marked with color.

<table>
<thead>
<tr>
<th>Method</th>
<th>Hip+</th>
<th>Knee+</th>
<th>Ankle+</th>
<th>Torso</th>
<th>Neck</th>
<th>Head</th>
<th>Nose</th>
<th>Shoulder+</th>
<th>Elbow+</th>
<th>Wrist+</th>
<th>X</th>
<th>Y</th>
<th>Z (Depth)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline-1*: 24.6</td>
<td>49.0</td>
<td>73.8</td>
<td>40.6</td>
<td>51.9</td>
<td>55.6</td>
<td>56.9</td>
<td>52.5</td>
<td>66.8</td>
<td>84.8</td>
<td>14.6</td>
<td>19.4</td>
<td>39.8</td>
<td></td>
</tr>
<tr>
<td>Baseline-2**: 23.5</td>
<td>11</td>
<td>49.7</td>
<td>10.7</td>
<td>72.6</td>
<td>12.2</td>
<td>36.8</td>
<td>3.8</td>
<td>50.4</td>
<td>1.5</td>
<td>53.0</td>
<td>2.6</td>
<td>49.6</td>
<td>7.3</td>
</tr>
<tr>
<td>Ours: 21.6</td>
<td>-3.0</td>
<td>45.7</td>
<td>-3.3</td>
<td>68.9</td>
<td>-4.9</td>
<td>35.2</td>
<td>-5.4</td>
<td>47.9</td>
<td>-4.0</td>
<td>51.1</td>
<td>-4.5</td>
<td>52.3</td>
<td>-4.6</td>
</tr>
</tbody>
</table>

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 right joint.

5. Conclusion

We introduced a new COCO-based large-scale, high-precision dataset for human body orientation estimation in 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 the 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 to use our dataset for other vision tasks, e.g., Person ReID.

Acknowledgements

Part computing work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. J.Z. Wang was supported by NSF grant no. 1921783.
References

[1] https://pytorch.org


CVPR Supplementary Material

A. Labeling Tool

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

![Figure A1. Labeling tool user interface.](image)

B. More Details on the HBOE Experiments

![Figure A2. Breakdown analysis of the performance of our HBOE baseline model.](image)

**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 trained on the TUD dataset respectively, using 1) the test set of the MEBow dataset (Fig. A2(a)); 2) the test set of the TUD dataset (Fig. A2(b)). Based on the same set of experiments for 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 generalization capability than it trained on TUD dataset. Second, we show how our baseline HBOE model performs when the camera point of view is the Front, Back, Left, and Right breakdown categories are shown in Fig. A2(c). It’s not surprising that our model performs best when the camera point of view is towards the Front of the person since 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 the 3-D human pose experiments with Protocol I in [47], the evaluation result is shown as in Table A1.

<table>
<thead>
<tr>
<th>Method</th>
<th>PA MPJPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chen et al. [11]</td>
<td>82.7</td>
</tr>
<tr>
<td>Moreno et al. [31]</td>
<td>76.5</td>
</tr>
<tr>
<td>Zhou et al. [54]</td>
<td>55.3</td>
</tr>
<tr>
<td>Sun et al. [46]</td>
<td>48.3</td>
</tr>
<tr>
<td>Sharma et al. [43]</td>
<td>40.9</td>
</tr>
<tr>
<td>Sun et al. [47]</td>
<td>40.6</td>
</tr>
<tr>
<td>Moon et al. [30]</td>
<td>34.0</td>
</tr>
<tr>
<td>Baseline*</td>
<td>34.9</td>
</tr>
<tr>
<td>Baseline 2**</td>
<td>34.4</td>
</tr>
<tr>
<td>ours</td>
<td>33.4</td>
</tr>
</tbody>
</table>

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. [47], trained on Human3.6M + MPII, as in the original paper. **Our baseline 2 is a re-implementation of Sun et al. [47], trained on Human3.6M + MPII + COCO (2-D Pose).

More Qualitative Human Body Orientation Estimation Examples are displayed in Fig. A3 to supplement the Fig. 4 in the main paper.

D. 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 displayed in Fig. A4.

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