Towards Fine-Grained HBOE with Rendered Orientation Set and Laplace Smoothing

Ruisi Zhao[1,2], Mingming Li[1], Zheng Yang[2], Binbin Lin[3], Xiaohui Zhong[4], Xiaobo Ren[4], Deng Cai[1,2], Boxi Wu[3] *

[1]State Key Lab of CAD&CG, Zhejiang University
[2]FABU Inc
[3]School of Software Technology, Zhejiang University
[4]Ningbo Zhoushan Port Group Co.,Ltd., Ningbo, China
zhaors00@zju.edu.cn

Abstract

Human body orientation estimation (HBOE) aims to estimate the orientation of a human body relative to the camera's frontal view. Despite recent advancements in this field, there still exist limitations in achieving fine-grained results. We identify certain defects and propose corresponding approaches as follows: 1). Existing datasets suffer from non-uniform angle distributions, resulting in sparse image data for certain angles. To provide comprehensive and high-quality data, we introduce *RMOS (**R**endered **M**odel **O**rientation **S**et)*, a rendered dataset comprising 150K accurately labeled human instances with a wide range of orientations. 2). Directly using one-hot vector as labels may overlook the similarity between angle labels, leading to poor supervision. And converting the predictions from radians to degrees enlarges the regression error. To enhance supervision, we employ Laplace smoothing to vectorize the label, which contains more information. For fine-grained predictions, we adopt weighted Smooth-L1-loss to align predictions with the smoothed-label, thus providing robust supervision. 3). Previous works ignore body-part-specific information, resulting in coarse predictions. By employing local-window self-attention, our model could utilize different body part information for more precise orientation estimations. We validate the effectiveness of our method in the benchmarks with extensive experiments and show that our method outperforms state-of-the-art. Project is available at: https://github.com/Whalesong-zrs/Towards-Fine-grained-HBOE.

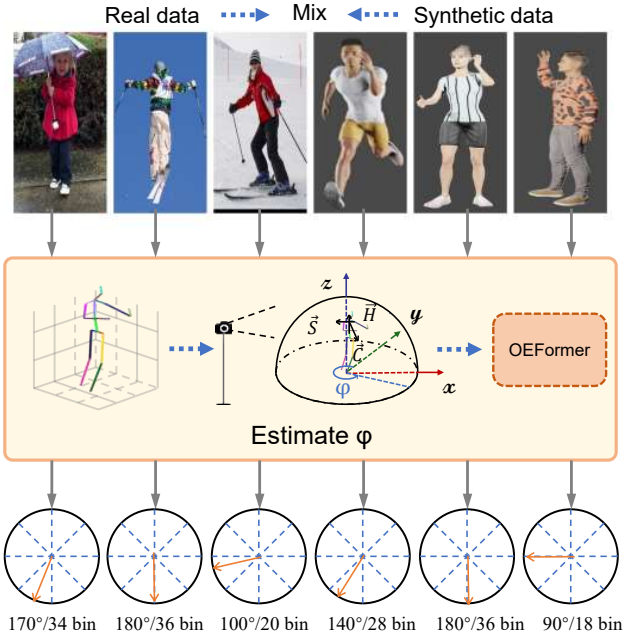

Figure 1: Illustration of HBOE. We present the real and synthetic data as examples, where the ϕ is the angle we estimating. We visualize the angles in the bottom lines.

Introduction

Human body orientation estimation (HBOE) involves estimating the orientation of a person's skeleton relative to the camera frontal view. It has been applied in various industrial applications, *e.g.,* pedestrian trajectory prediction in autonomous driving and human-robot interactions. For some downstream tasks, body orientation is easier to obtain, provides sufficient information, and demonstrates greater robustness to illumination and occlusions compared to 3D pose estimation, for understanding human behaviors.

Though many prior methods performed well in coarse-grained estimation, they encountered challenges in accurately predicting angles. These difficulties can be attributed to three key factors. The primary bottleneck is the quality of datasets. The widely used TUD dataset (Andriluka, Roth, and Schiele 2010), which originally only provided 8-class labels, has a small scale that limits model capability. Additionally, the high-quality MEBOW (Wu et al. 2020) dataset based on COCO (Lin et al. 2014) suffers from a non-uniform angle distribution and scarce image data for certain angles, causing inaccurate predictions. Secondly, many methods treated HBOE as a 6/8-class classification problem (Liu, Liu, and Ma 2017; Yu et al. 2019; Choi, Lee, and Zhang 2016), further contributing to coarse results and ignoring the similarity between classes. Directly regressing values in radians struggles to obtain accurate predictions, as converting radians to degrees enlarges the regression error (Zhou et al. 2022; Burgermeister and Curio 2022). More-

*Corresponding author.

over, people usually judge the fine-grained HBOE by observing the shape of body core parts. Unfortunately, previous works tended to overlook the prior of human perception in fine-grained HBOE and adopted shallow model architectures (Raza et al. 2018; Choi, Lee, and Zhang 2016; Zhou et al. 2022; Burgermeister and Curio 2022), resulting in underfitting and reduced performance.

To address limitations of existing datasets, we present the RMOS (Rendered Model Orientation Set), a synthetic dataset containing 150K images with detailed annotations. To achieve high-precision labels, we divide 360° into 72 bins, while ensuring that multiple images are rendered for each bin to guarantee uniform angle coverage. To promote diversity, our dataset captures 10 differently dressed models in 48 daily poses i.e., *running, standing arms out, yoga pose*, from five different viewpoints. We highlight the advantages of incorporating synthetic data during training.

Considering the correlation between angle labels, we employ a Laplace smoothing strategy inspired by previous works (Müller, Kornblith, and Hinton 2019; Wu et al. 2020). We assign the highest probability to the label's corresponding bin while ensuring certain probabilities for adjacent bins using a Laplacian kernel. During the training process, we also adopt the weighted Smooth-L1-loss, aligning the predictions with the smoothed-labels. Compared to Gaussian smoothing, our method's peak probability is higher, which enhances class information distinction. By incorporating this strategy, our method significantly outperforms existing techniques, achieving more precise orientation estimations.

Based on observations of human perception for HBOE, we adopt the local-window self-attention, which partitions the feature map and individually applies multi-head self-attention to each segment. Furthermore, to tackle the challenges in HBOE, we propose Orientation Estimating Former (OEFormer) based on HRFormer (Yuan et al. 2021). Compared to the vanilla HRFormer, we employ a deeper network architecture with multiple early-stage branches for more comprehensive feature extraction. Considering the varying resolutions of human images in the original data, effective feature extraction has become crucial. In the final stage, we integrate feature maps generated from different stages to achieve a more accurate final prediction.

Our main contributions in this work are:

1. We introduce a novel rendered dataset that provides high-precision and comprehensive orientation annotations, effectively addressing the gaps in existing datasets.

2. Employing a Laplace smoothing strategy and weighted Smooth-L1-loss, we enhance the alignment between ground truth and predictions, resulting in more effective training and markedly improved accuracy in orientation estimation.

3. This is the first application of a transformer-based model for HBOE. We present a powerful model, OEFormer, which involves local-window self-attention to focus on body-part-specific information and outperforms state-of-the-art in existing benchmarks.

Related Works

Body orientation estimations methods. Classical studies in HBOE primarily relied on feature engineering and classifiers like HOG/linSVM (Flohr et al. 2014), limited by dataset quality and scale. Enzweiler *et al.* (Enzweiler and Gavrila 2010) classified the pedestrian and used Gaussian mixture model to estimate orientation. Previous deep learning methods also treated this task as a classification problem, with various approaches like 4-layer neural networks (Choi, Lee, and Zhang 2016) and 14-layer convolutional networks (Raza et al. 2018). To get fine-grained predictions, the method in (Yu et al. 2019) leveraged keypoint detection results from another 2D pose estimation model as an auxiliary condition. The TUD multiview pedestrians dataset (Andriluka, Roth, and Schiele 2010) has long served as a key benchmark in HBOE. It was further improved by Hara et al. (Hara and Chellappa 2017) who relabeled it with continuous annotations, facilitating extensive use in early research. With the appearance of MEBOW (Wu et al. 2020), TUD evolved into a test benchmark to evaluate model generalizability. MEBOW, the largest benchmark in this field, offers high-precision annotations and varied backgrounds in real-world settings, containing 130K human instances. PedRecNet (Burgermeister and Curio 2022) utilized this benchmark to address orientation estimation challenges, whereas JOINT-Net (Zhou et al. 2022) first detected human instances and then estimated their orientation.

Synthetic Data for Image Recognition. In computer vision tasks, utilizing synthetic data for augmentation is a widely adopted strategy. Previous works often employed rendered 2D instances or 3D model scene data with graphics engines (Dosovitskiy et al. 2015; Peng et al. 2017; Richter et al. 2016). In HBOE, PedRecNet (Burgermeister and Curio 2022) also utilized synthetic data to increase data diversity, however, it faced certain performance limitations. Generative models have recently gained popularity (Ho, Jain, and Abbeel 2020; Besnier et al. 2020). These methods leverage generated data to solve various vision tasks, including classification (Azizi et al. 2023), semantic segmentation (Zhang et al. 2021), and contrastive learning (Jahanian et al. 2021). However, these models may struggle with accurately capturing human orientation information.

Attention Mechanism. The success of self-attention (Vaswani et al. 2017) has opened new avenues for exploring attention strategies in deep learning. Recently, attention mechanisms have been applied to various visual tasks, *e.g.,* image classification (Hu, Shen, and Sun 2018), object detection (Dai et al. 2017), semantic segmentation (Fu et al. 2019) and pose estimation (Chu et al. 2017). As a pivotal type of attention mechanism, local attention has played a significant role in numerous works. SASA (Ramachandran et al. 2019) suggests that using self-attention to gather global information can be computationally intensive. The authors demonstrate that local attention not only improves computational efficiency but also elevates the quality of results. Concurrent with SASA, LR-Net (Hu et al. 2019) employed local attention for image classification. Similarly, HRFormer (Yuan et al. 2021) utilized local self-attention for various vision tasks, showing its versatility. In this work, we adopt an attention mechanism akin to these previous studies, employing local-window self-attention to specifically focus on different body part information for more accurate estimations.

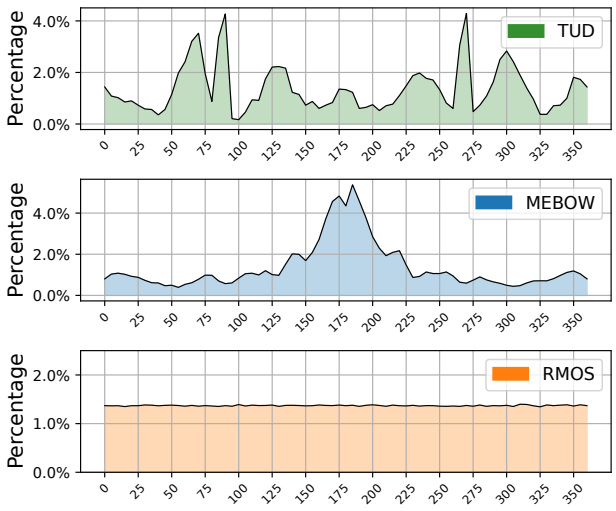

Figure 2: Distributions of datasets. The x-axis represents the orientation labels and the y-axis represents the corresponding percentage. Our RMOS has a uniform data distribution.

Methodology

In this section, we first introduce the definition of body orientation. Next we present our innovative approach to creating the RMOS. Additionally, we describe the OEFormer design. Finally we present the Laplace smoothing and weighted Smooth-L1-loss adopted to HBOE.

Definition of Body Orientation

As shown in the Fig. 1, we consider the camera's shooting direction as the positive **y**-axis direction and the image plane as z-x plane. The orientation ϕ we aim to determine is the angle between the projection of the chest facing direction onto the x-y plane and the y-axis. Given a human pose, the chest facing direction $\overrightarrow{C}$ can be denoted as $\overrightarrow{C} = \overrightarrow{S} \times \overrightarrow{H}$, where $\overrightarrow{S}$ is the direction from left shoulder to right shoulder, and $\overrightarrow{H}$ is the direction from hip to neck. Given the projection $\overrightarrow{C_{xy}}$ of $\overrightarrow{C}$, and y-axis direction $\overrightarrow{O_y}$, the ϕ can be compute as:

$$\cos\phi = \frac{\overrightarrow{C_{xy}} \cdot \overrightarrow{Oy}}{\|\overrightarrow{C_{xy}}\|\|\overrightarrow{Oy}\|} . \tag{1}$$

The output probability distribution p is obtained by passing image x through model f and followed by a softmax function. The label y is a 72-dimensional one-hot vector, where the element at index i is set to 1. The index i is:

$$i = \text{round}(\frac{\phi}{5}),\ \phi \in [0°, 360°). \tag{2}$$

Dataset Creation

In this work, we propose augmenting training data with synthetic data to enhance the model training process. To facilitate this, we choose Blender [1] for modeling and rendering for the following reasons: 1) Blender supports Python scripting to control camera movements through commands, significantly simplifying the image capturing process. 2) It allows easy modifications to the model's shape and appearance, providing diverse transformations to diversify the dataset. During rendering, we keep the model's position fixed while rotating the camera around it, capturing images every 5-° rotation to obtain fine-grained labels. For modeling, we manipulate the model's skeleton to achieve pose variations.

To maintain diversity, RMOS includes 10 differently dressed models, each capable of 48 common daily poses. We capture data from five shooting views i.e., *downward shot, overhead shot*, which correspond to different shooting perspectives encountered in real-world scenarios. Fig. 2 illustrates the angle distribution of existing datasets. Evidently, RMOS encompasses all angles while containing substantial image data for each one. Furthermore, RMOS benefits from non-overlapping human instances, reducing potential model misinterpretations.

Model Architecture

People often judge orientation based on the different body part appearances. Therefore, we enhance attention to different body regions using local-window self-attention. In real-world scenarios, human body instance clarity varies, making precise estimation challenging. To address this, we employ network branches with different resolutions to gather sufficient information, which is then consolidated to obtain a comprehensive result.

Following the multi-resolution parallel design (Wang et al. 2020; Yuan et al. 2021), we present our OEFomer architecture in Fig. 3. As (Dai et al. 2021; Xiao et al. 2021) suggested, we utilize convolutional layer in both the stem and the first stage to extract feature. Subsequently, transformer blocks with local-window self-attention are employed in later stages. Our architecture comprises various branches with different resolutions in each stage, ranging from high to low. Upsampling and downsampling operations enable information exchange between stages for effective feature extraction and fusion. In contrast to HRformer, we employ more modules and additional branches in the first stage to enable earlier and more comprehensive feature extraction, aiming to improve overall performance . Following these stages, outputs from different stages are concatenated into residual blocks (He et al. 2016) to obtain the final result.

In local-window self-attention, we divide the feature map $X \in \mathbb{R}^{N \times D}$ into M partitions, where each partition has a size K × K. We perform multi-head self-attention in each partition m. In this work K is set to 7. The representation of local self-attention as in Fig. 4.

Loss Function

In fine-grained HBOE tasks, directly mapping labels to one-hot vectors overlooks similarities between adjacent labels. For example, 355° and 0° are highly similar but have low similarity as one-hot vectors. When the predicted category of the model is close but not equal to the ground truth, the supervision effect of the cross-entropy is poor. While regression provides some supervision, it still falls short of high

[1] https://www.blender.org/

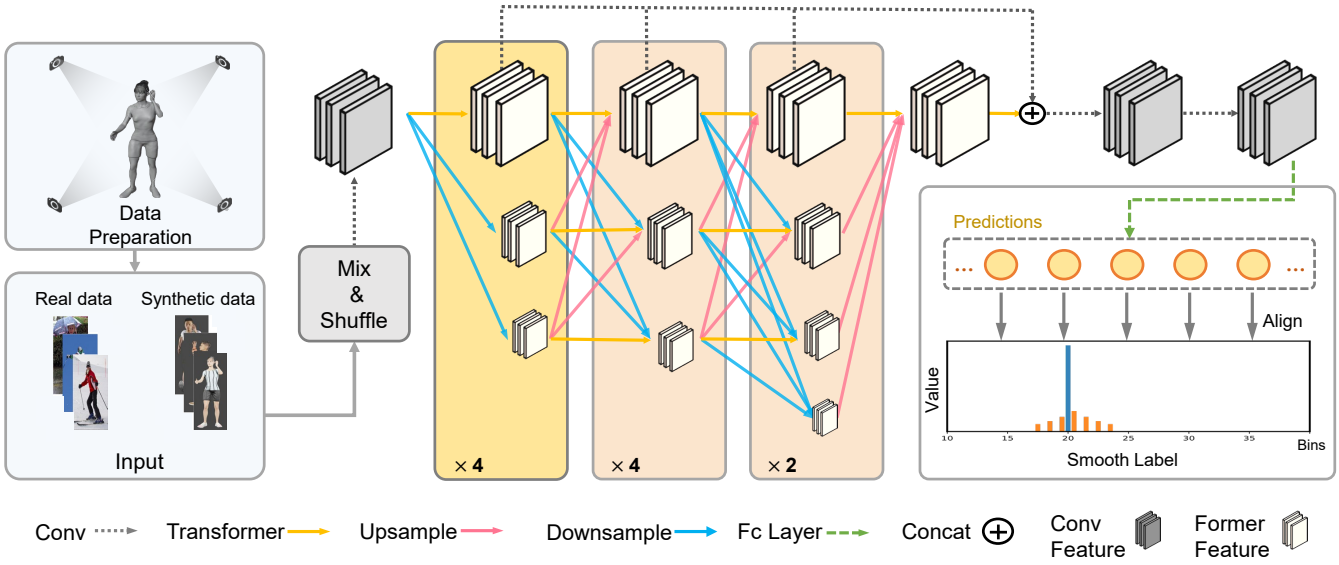

Figure 3: Illustrating the OEFormer architecture. We train on combined real and synthetic data. Our focused improvements (in yellow) utilize more modules and branches with increased depth to enhance early feature extraction. Predictions are generated by fusing features from different stages and aligned with smoothed labels.

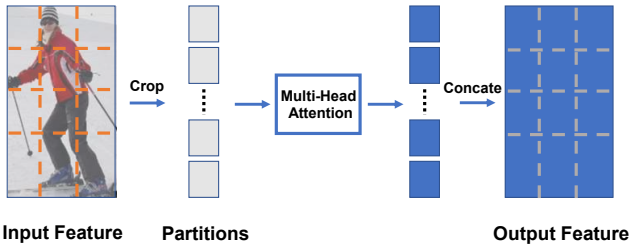

Figure 4: Local-window self-attention for HBOE.

precision. For instance, in cosine regression, as the difference approaches 0.1, the angle error nears 11.5°, leading to coarse prediction.

Considering these supervision limitations and inspired by (Müller, Kornblith, and Hinton 2019; Wu et al. 2020), we perform Laplace smoothing to maximize the label peak probability while maintaining certainty for adjacent categories, aligning with human intuition. Here y is a one-hot vector of length 72. The Laplace kernel K_l on y_i:

$$K_l(y_i) = \frac{1}{2\sigma} e^{-\frac{|y_i|}{\sigma}} . \tag{3}$$

The window size of the Laplace kernel w is set to 9. The smoothed label $\hat{y}$ after Laplace smoothing can be obtained as follows:

$$\hat{y}_i = (y * K_l)(i) = \sum_{j \in S_i} y_j K_l(y_j) , \tag{4}$$

$$S_i = \left\{ (i+k) \bmod 72 \mid \ k = \lfloor -\frac{w}{2} \rfloor, \ldots, \lfloor \frac{w}{2} + 1 \rfloor \right\} . \tag{5}$$

Given the model output p and the smoothed label $\hat{y}$, our loss function is:

$$\mathcal{L}(p, \hat{y}) = \begin{cases} 0.5 \times (p - \hat{y})^2 / \beta & \text{if } |p - \hat{y}| < \beta, \\ |p - \hat{y}| - 0.5 \times \beta & otherwise. \end{cases} \tag{6}$$

We use this weighted Smooth-L1-Loss to align p with the $\hat{y}$, and the β is the weight. This approach allows for better capturing of angle errors present in the real world and more accurately expressing the model's confidence in different orientation bins.

Experiments

In this section, we compare performance of different backbones for HBOE and demonstrate OEFormer's superiority over other models. Next, we compare various supervised methods including our weighted Smooth-L1-loss, Wu *et al.*'s gaussian mapping loss, cross-entropy, and cosine regression. Additionally, we try to upsample data of rare orientations instead of adding RMOS, however, the results are not as expected. Ablation experiments investigate the impact of different RMOS proportions and σ in Laplace kernel on performance. Furthermore, we compare with existing methods and find that incorporating RMOS leads to better results in fine-grained metrics and MAE. Through these experiments, we comprehensively evaluate the performance and stability of our proposed HBOE method.

Experimental Setup

Datasets. As the largest and most valuable real-scene dataset, the MEBOW dataset contains around 130K training samples and has rich background environments. It will be

Backbone	Training	Train set	Test set	*Acc.*(5°)↑	*Acc.*(15°)↑	*Acc.*(22.5°)↑	*Acc.*(30°)↑	*Acc.*(45°)↑	MAE(°)↓
ResNet50	From-scratch	MEBOW	MEBOW	61.1	83.9	88.5	92.3	94.5	13.11
HRNet				63.0	85.2	89.3	92.8	95.2	12.892
HRFormer				62.4	85.5	89.2	92,7	94.9	12.473
OEFormer				63.1	85.4	89.4	93.1	**95.2**	12.458
HRNet		RMOS + MEBOW		63.4	85.7	**89.3**	92.9	95.1	12.831
OEFormer				**64.4**	**85.8**	**89.3**	**93.2**	95.1	**12.135**
ResNet50	Fine-tune	MEBOW	MEBOW	67.3	89.1	92.9	95.9	97.5	9.200
HRNet				68.4	90.8	93.6	96.5	**97.9**	8.479
HRFormer				69.9	90.6	93.9	96.6	97.7	9.344
OEFormer				70.6	90.5	93.6	96.5	97.8	8.400
HRNet		RMOS +MEBOW		70.8	90.8	93.6	**96.7**	**97.9**	8.384
OEFormer				**72.1**	**91.0**	**94.0**	96.6	**97.9**	**8.129**

Table 1: Performance comparison of different backbones in the HBOE task, including different choices of models, training schedules, training data. ↑ indicates that, as the metric improves, the performance improves. ↓ indicates that, as the metric decreases, the performance improves.

used for both training and testing. Additionally, we will incorporate the RMOS dataset as supplementary training data and evaluate its value on the MEBOW test set. We don't use any special training techniques and are able to achieve good results by simply mixing and training the RMOS and MEBOW together. The data in the TUD dataset has clear and complete human body shapes and provides continuous labels. Due to the relatively small scale of the TUD dataset, following previous methods, we will train on MEBOW and test on TUD to assess generality.

Evaluation metrics. As in previous works, we adopt Accuracy-22.5°, Accuracy-45° and mean absolute error (MAE) as evaluation metrics. Accuracy-X° represents the percentage of predictions within X° of ground truth, while MAE evaluates overall performance. Following previous work (Wu et al. 2020) and leveraging the precise labels provided by MEBOW and RMOS, we include Accuracy-5°, Accuracy-15° and Accuracy-30° in our evaluation analysis.

Training Protocol. Input instances are cropped and resized to 256×192 while applying data augmentation techniques including flipping and scaling. For OEFormer training, we use 80 epochs with a batch size of 256 and the AdamW optimizer with initial learning rate 1×10^{-5}. We set β to 0.2 and σ to 2.0 for the loss function. For the experiments in Tab. 1, we implenment these backbones based on the mmpose (Contributors 2020). And all these experiments used the same set.

Fine-Grained HBOE Performance

We conducted a performance comparison of various commonly used backbones for HBOE tasks, including ResNet and the HRNet used by Wu *et al.*, which achieved promising results. Additionally, we also compared our OEFormer with HRFormer to demonstrate its excellent performance in fine-grained results. As mentioned in Wu *et al.*, using pretrained models based on 2D pose estimation can effectively improve model performance. Therefore, we conducted two groups of experiments: one trained from scratch, and one fine-tuned using a pretrained model for 2D pose estimation.

As shown in the Tab. 1, using pretrained models yields better results than training from scratch. Regardless of training approach, our model outperforms others in effectiveness. Although the HRFormer architecture achieves good accuracy, it falls short in MAE. By utilizing attention mechanisms, our model surpasses Wu *et al.* previous best method, achieving improved fine-grained accuracy and lower MAE. Compared to other transformer models, we also achieve superior results by extracting more features earlier.

We implemented and compared five supervised methods: our weighted Smooth-L1-loss, gaussian-mapping loss (Wu et al. 2020), cross-entropy loss, and cosine regression loss. As shown in Tab. 3, direct regression performs reasonably for coarse-grained evaluation metric. Treating HBOE as a 72-class classification task and neglecting the similarity between labels. When the model assigns higher probabilities to categories close to the label, it indicates that the model has some capability in assessing angle information. However, using standard cross-entropy loss can still result in large losses, thus introducing training bias. Our loss function takes this into consideration and achieves good overall results. Compared to Wu's smoothing method, our method ensures higher peak probability values on label classes, and achieves better results in fine-grained tasks. We set experiment to assign higher weights to rare sample categories in the loss function. However, it faces performance limitations.

Abalation Studies

Analysis of RMOS. In this part, we investigate the impact of different proportions of the RMOS dataset on model performance. Specifically, we utilize MEBOW-Net and conduct experiments incorporating the MEBOW dataset for training. We progressively introduce 20%, 50%, 70%, and 100% of the RMOS data, while keeping the MEBOW data constant. The experimental results are presented in Tab. 4.

The ablation studies demonstrate that incorporating the RMOS leads to consistent improvements in model performance across all metrics. As we increase the percentage of RMOS data from 20% to 100%, both fine-grained and coarse-grained accuracy steadily improve. This indicates that our synthesized RMOS data complements the real-

Method	Train set	Test set	*Acc.*(5°) ↑	*Acc.*(15°) ↑	*Acc.*(22.5°) ↑	*Acc.*(30°) ↑	*Acc.*(45°) ↑	MAE(°) ↓
MEBOW-Net (2020)	MEBOW	MEBOW	68.6	90.7	93.9	**96.9**	**98.2**	8.393
Joint-Net (2022)			48.3	85.2	91.0	93.2	96.5	10.526
PedRecNet (2022)			52.0	86.2	92.3	95.1	97.0	9.702
ours			71.1	90.5	93.6	96.5	97.8	8.356
MEBOW-Net (2020)	RMOS +MEBOW	MEBOW	70.8	**91.0**	93.6	96.7	97.9	8.384
ours			**72.1**	**91.0**	**94.0**	96.6	97.9	**8.129**
Hara (2017)	TUD	TUD	-	-	70.6	-	86.1	26.6
AKRF-VW (2018)			-	-	68.6	-	78	34.7
Yu (2019)			-	-	75.7	-	96.8	15.3
MEBOW-Net (2020)	MEBOW	TUD	39.5	66.7	77.3	92.2	99.0	14.191
PedRecNet (2022)			-	-	79.6	-	99.0	13.702
ours			**41.7**	**72.5**	**83.2**	**95.5**	**99.7**	**12.298**

Table 2: Performance comparison of existing methods in the HBOE task. The column Train set specifies the training dataset(s) used to train the models. Test set specifies on which test sets the results are reported on.

Supervised meethod	MAE	*Acc.*(5°)	*Acc.*(30°)	*Acc.*(45°)
Smooth-L1-loss	**8.129**	**72.1**	**96.6**	**97.9**
Wu *et al.*	8.333	71.0	96.4	97.8
cross-entropy	21.508	44.5	83.8	88.4
cosine-regression	16.772	31.0	87.0	92.5
Re-weight loss	9.332	65.3	96.1	97.4

Table 3: Comparison between different supervised methods.

world data distribution in MEBOW, providing valuable additional training examples that enhance the model's estimation capabilities. The optimal performance is achieved when utilizing the full RMOS dataset, suggesting it provides comprehensive coverage of body orientations. Our experiments provide insights into the benefits of supplementing high-quality synthetic data for advancing HBOE.

We validated the model's generalization performance on the real-world dataset TUD. As shown in Tab. 5, the domain gap between synthetic data and real-world data has not affected the model's performance, and the supplementation of data distribution has improved the model's generalization.

Proportions	*Acc.*(5°)	*Acc.*(30°)	*Acc.*(45°)	MAE
0	68.4	96.5	97.9	8.479
20 %	69.9	96.7	97.9	8.483
50 %	70.2	96.5	97.8	8.447
70 %	70.4	96.5	97.8	8.407
100 %	**70.8**	**96.7**	**97.7**	**8.384**

Table 4: Ablation study on the addition of RMOS. Experiment is done on MEBOW-Net.

Analysis of σ in Laplace kernel. The σ value in label smoothing affects the shape of the predicted probability distributions. Smaller σ values concentrate more probability mass on the ground-truth class, resulting in sharper peaks in the distributions. This compels the model to make highly confident predictions focused on the true label. In our experiments, we evaluate models trained with various σ values using fine-grained accuracy metrics that reward correct classification, and coarse-grained metrics that measure generalization capability.

Dataset	*Acc.*5°	*Acc.*15°	*Acc.*22.5°
w/o RMOS	**39.5**	66.7	77.3
w RMOS	36.6	**67.7**	**78.0**

Table 5: Evaluation on TUD Dataset

σ	*Acc.*(5°)	*Acc.*(22.5°)	*Acc.*(45°)	MAE
1.0	71.0	93.6	97.2	8.308
2.0	**72.1**	**94.0**	97.9	**8.129**
3.0	71.2	93.8	97.8	8.446
4.0	69.7	**94.0**	**98.0**	8.408

Table 6: Ablation study on σ.

We find that small σ values like 1.0 yield superior fine-grained performance. Larger σ values of 3.0 and 4.0 enhance coarse-grained performance by diffusing probability, although at the cost of lower fine-grained accuracy. By balancing these factors, we select σ=2.0, which provides confident prediction while retaining generalizability.

Comparison of Different Methods

In order to comprehensively evaluate the performance of our proposed method, we conducted extensive comparisons against the previous state-of-the-art techniques for HBOE. As shown in Tab. 2, we compare our method with MEBOW-Net (Wu et al. 2020), Joint-Net (Zhou et al. 2022) and PedrecNet (Burgermeister and Curio 2022). Without any additional synthetic data, our method was able to surpass all existing state-of-the-art methods on MEBOW in terms of fine-grained accuracy metrics and mean absolute error. This demonstrates the strengths of our approach even when trained solely on real-world data.

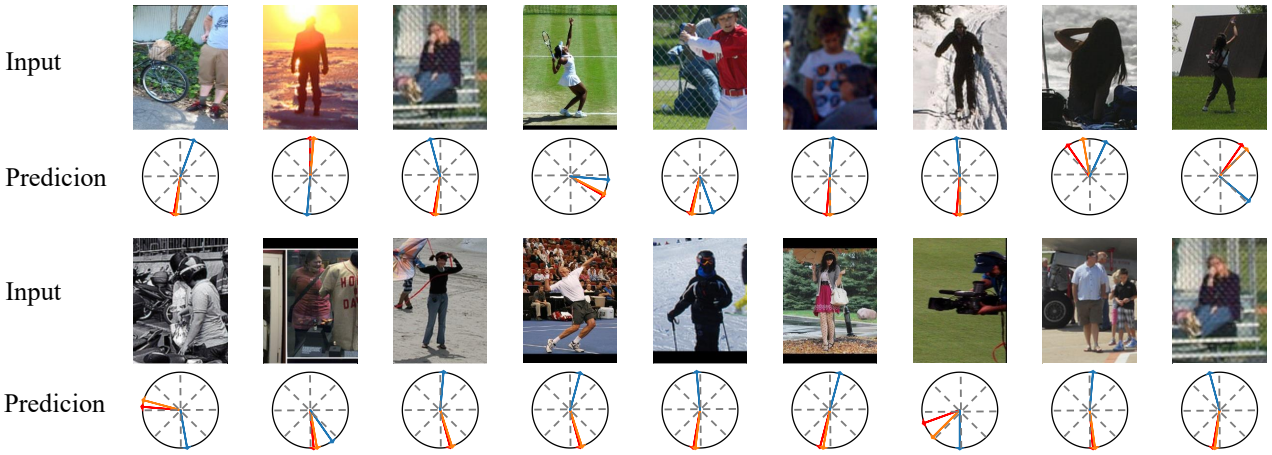

Figure 5: HBOE results generated by OEFomer (trained on MEBOW and RMOS, σ=2), ● : ground truth, ● : our prediction, ● : Wu *et.al* prediction. It can be observed that our method is able to accurately determine the orientation of the human body even in cases where body are occluded or the image resolution is low.

Dataset	Angle	Train data	Test data	*Acc.*5°
w/o RMOS	30°	942	45	60.0
w RMOS				**75.6**
w/o RMOS	205°	2951	117	76.1
w RMOS				**83.8**

Table 7: Evaluations for some categories in MEBOW

We then incorporated our novel RMOS synthetic data into the training process of both our method and a strong baseline model MEBOW-Net. The results clearly validated the benefits of RMOS. With this additional synthetic data, both methods exhibited significant improvements on fine-grained evaluation metrics and achieved lower MAE, highlighting the usefulness of our proposed data augmentation technique. And we test our method's generalizability on TUD.

As shown in Tab. 7, the integration of RMOS contributes to an increase in the model's fine-grained discrimination ability, both for categories with limited training data and those with abundant training data.

In Fig. 5, we demonstrate the performance comparison between our method and MEBOW-Net. In these examples, only partial human bodies appear in the images. With our local-window self-attention, our method can make more accurate estimations for such cases. Due to factors like poor illumination and low resolution, previous method fails to correctly judge the front/back side of the human bodies, making predictions opposite to the labels. In contrast, our method can generate correct predictions. When human bodies make large-scale motions, our method exhibits good robustness.

In Figure 6, we enhance our analysis by overlaying heatmaps onto the original images, thereby visualizing the final feature maps of our model. This technique effectively demonstrates the specific areas within the images that our model focuses on. Notably, the heatmaps reveal a significant

Figure 6: Heatmap of different methods

concentration of the model's attention on the core regions of the human body. This pattern of focus is in harmony with the general principles of human perception, where central body parts are often crucial for interpreting posture and actions. The alignment of our model's focus with these perceptual norms underscores its potential applicability in fields that require an in-depth understanding of human body dynamics.

Conclusion

In this paper, we comprehensivily analyzed the underlying factors behind the poor performance of existing methods in fine-grained HBOE tasks. These reasons include non-uniform distribution of existing datasets, inadequate supervisory approaches, and the relatively simplistic model architectures used in previous methods. Consequently, our primary proposition involves augmenting real data deficits by introducing synthetic data for data augmentation. Furthermore, we employ a label smoothing strategy to transform original angle labels into smoothed vectors, incorporating a weighted Smooth-L1-loss for effective supervision. Lastly, we adopted local-window self-attention mechanism, presenting a transformer-based model architecture that yields remarkable performance enhancements.

Acknowledgments

This work was supported in part by The National Nature Science Foundation of China (Grant Nos: 62273302, 62036009, 61936006, 62273303), in part by Yongjiang Talent Introduction Programme (Grant No: 2023A-197-G).

References

Andriluka, M.; Roth, S.; and Schiele, B. 2010. Monocular 3d pose estimation and tracking by detection. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 623–630. Ieee.

Azizi, S.; Kornblith, S.; Saharia, C.; Norouzi, M.; and Fleet, D. J. 2023. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*.

Besnier, V.; Jain, H.; Bursuc, A.; Cord, M.; and Pérez, P. 2020. This dataset does not exist: training models from generated images. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Burgermeister, D.; and Curio, C. 2022. PedRecNet: Multi-task deep neural network for full 3D human pose and orientation estimation. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, 441–448. IEEE.

Choi, J.; Lee, B.-J.; and Zhang, B.-T. 2016. Human body orientation estimation using convolutional neural network. *arXiv preprint arXiv:1609.01984*.

Chu, X.; Yang, W.; Ouyang, W.; Ma, C.; Yuille, A. L.; and Wang, X. 2017. Multi-context attention for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1831–1840.

Contributors, M. 2020. OpenMMLab Pose Estimation Toolbox and Benchmark. https://github.com/open-mmlab/mmpose.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 764–773.

Dai, Z.; Liu, H.; Le, Q. V.; and Tan, M. 2021. Coatnet: Marrying convolution and attention for all data sizes. *Advances in neural information processing systems*, 34: 3965–3977.

Dosovitskiy, A.; Fischer, P.; Ilg, E.; Hausser, P.; Hazirbas, C.; Golkov, V.; van der Smagt, P.; Cremers, D.; and Brox, T. 2015. FlowNet: Learning Optical Flow With Convolutional Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Enzweiler, M.; and Gavrila, D. M. 2010. Integrated pedestrian classification and orientation estimation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 982–989. IEEE.

Flohr, F.; Dumitru-Guzu, M.; Kooij, J. F. P.; and Gavrila, D. M. 2014. Joint probabilistic pedestrian head and body orientation estimation. In *2014 IEEE Intelligent Vehicles Symposium Proceedings*, 617–622.

Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; and Lu, H. 2019. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3146–3154.

Hara, K.; and Chellappa, R. 2017. Growing regression tree forests by classification for continuous object pose estimation. *International Journal of Computer Vision*, 122: 292–312.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, H.; Zhang, Z.; Xie, Z.; and Lin, S. 2019. Local relation networks for image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3464–3473.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Jahanian, A.; Puig, X.; Tian, Y.; and Isola, P. 2021. Generative models as a data source for multiview representation learning. *arXiv preprint arXiv:2106.05258*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, P.; Liu, W.; and Ma, H. 2017. Weighted sequence loss based spatial-temporal deep learning framework for human body orientation estimation. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, 97–102.

Müller, R.; Kornblith, S.; and Hinton, G. E. 2019. When does label smoothing help? *Advances in neural information processing systems*, 32.

Peng, X.; Usman, B.; Kaushik, N.; Hoffman, J.; Wang, D.; and Saenko, K. 2017. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*.

Ramachandran, P.; Parmar, N.; Vaswani, A.; Bello, I.; Levskaya, A.; and Shlens, J. 2019. Stand-alone self-attention in vision models. *Advances in neural information processing systems*, 32.

Raza, M.; Chen, Z.; Rehman, S.-U.; Wang, P.; and Bao, P. 2018. Appearance based pedestrians' head pose and body orientation estimation using deep learning. *Neurocomputing*, 272: 647–659.

Richter, S. R.; Vineet, V.; Roth, S.; and Koltun, V. 2016. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 102–118. Springer.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. 2020. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3349–3364.

Wu, C.; Chen, Y.; Luo, J.; Su, C.-C.; Dawane, A.; Hanzra, B.; Deng, Z.; Liu, B.; Wang, J. Z.; and Kuo, C.-h. 2020. MEBOW: Monocular estimation of body orientation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3451–3461.

Xiao, T.; Singh, M.; Mintun, E.; Darrell, T.; Dollár, P.; and Girshick, R. 2021. Early convolutions help transformers see better. *Advances in neural information processing systems*, 34: 30392–30400.

Yu, D.; Xiong, H.; Xu, Q.; Wang, J.; and Li, K. 2019. Continuous Pedestrian Orientation Estimation using Human Keypoints. In *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, 1–5.

Yuan, Y.; Fu, R.; Huang, L.; Lin, W.; Zhang, C.; Chen, X.; and Wang, J. 2021. Hrformer: High-resolution vision transformer for dense predict. *Advances in Neural Information Processing Systems*, 34: 7281–7293.

Zhang, Y.; Ling, H.; Gao, J.; Yin, K.; Lafleche, J.-F.; Barriuso, A.; Torralba, A.; and Fidler, S. 2021. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10145–10155.

Zhou, H.; Jiang, F.; Si, J.; and Lu, H. 2022. Joint Multi-Person Body Detection and Orientation Estimation via One Unified Embedding. *arXiv preprint arXiv:2210.15586*.