2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Human Orientation Estimation Under Partial Observation

Jieting Zhao[1], Hanjing Ye[1], Yu Zhan[1], Hao Luan[2] and Hong Zhang[1*]

***Abstract*— Reliable Human Orientation Estimation (HOE) from a monocular image is critical for autonomous agents to understand human intention. Significant progress has been made in HOE under full observation. However, the existing methods easily make a wrong prediction under partial observation and give it an unexpectedly high confidence. To solve the above problems, this study first develops a method called Part-HOE that estimates orientation from the visible joints of a target person so that it is able to handle partial observation. Subsequently, we introduce a confidence-aware orientation estimation method, enabling more accurate orientation estimation and reasonable confidence estimation under partial observation. The effectiveness of our method is validated on both public and custom-built datasets, and it shows great accuracy and reliability improvement in partial observation scenarios. In particular, we show in real experiments that our method can benefit the robustness and consistency of the Robot Person Following (RPF) task.**

I. INTRODUCTION

Human orientation, which in this context specifically refers to the human yaw angle, provides crucial information for many human-robot interaction applications such as Robot Person Following (RPF) [1]. RPF tasks like walking-aid robots [2], video filming drones [3], and autonomous trolleys [4], all require accurate and reliable human orientation information for the robot to calculate the desired following pose with respect to human.

In traditional RPF systems [4] [5], the human orientation is assumed to be aligned with the direction of movement. Human orientation can be obtained according to human velocity direction in a global frame. However, a consistent global frame needs an additional localization algorithm, which is not necessary for RPF systems. Even if global information is provided, traditional RPF systems still fail when the human is spinning without any positional change.

In contrast, Human Orientation Estimation (HOE) using monocular images does not rely on global position information. HOE has been researched for a long time in computer vision. Most early works extract handcraft features from images and estimate orientation using machine learning. Due to the constraints of the number of data and the capacity of the machine-learning model, early works show low accuracy and reliability. The development of deep neural networks (DNN) alongside human joint detection algorithms makes it possible to accurately estimate orientation. Some methods [6] [7] tried to leverage human joints as additional cues to the HOE task. Yu et al. [6] improved HOE accuracy a lot by utilizing deep networks to detect human joints and extract geometric features from the detected joints. This demonstrates that cues from human joints can play a crucial role in improving the accuracy of orientation estimation.

Despite the improvement of HOE accuracy, most orientation datasets built with a motion capture system are collected indoors with a clean background and contain no occlusion problems. As a result, algorithms developed by these datasets often struggle when confronted with partial observations, which are common in wheeled robots or robot dogs equipped with cameras. MEBOW [8] creates in-the-wild datasets and trains orientation estimation models concurrently with human joint detection. MEBOW improves the orientation estimation accuracy a lot by learning orientation in a regression manner and uses human joint detection to provide additional cues. Nevertheless, existing methods still struggle with partial observations because they have limited occlusion data and fail to recognize visible joints under partial observation, which are essential cues for the HOE task. In addition,

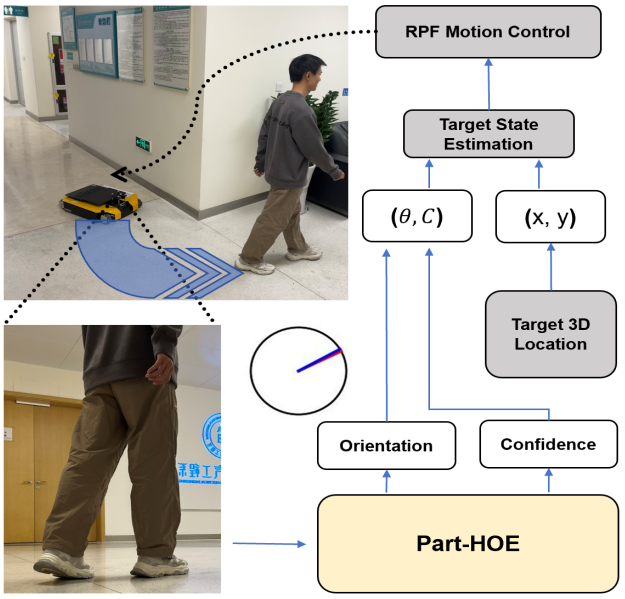


Fig. 1. Framework Overview: An example of partial observation in robot person following. The existing methods for Human Orientation Estimation (HOE) struggle in this scenario. We propose an occlusion-robust method Part-HOE utilizing visible joints to help target state estimation and to improve RPF performance.

*corresponding author (hzhang@sustech.edu.cn).
[1]Jieting Zhao, Hanjing Ye, Yu Zhan and Hong Zhang are with Shenzhen Key Laboratory of Robotics and Computer Vision, Southern University of Science and Technology (SUSTech), and the Department of Electronic and Electrical Engineering, SUSTech.
[2]Hao Luan is with the Department of Computer Science, School of Computing, National University of Singapore.
This work was supported by the Shenzhen Key Laboratory of Robotics and Computer Vision under Grant ZDSYS20220330160557001.
Code: https://github.com/zhaojieting/Part_HOE.

MEBOW regresses the orientation with a fixed Gaussian distribution, making the predicted orientation (represented by one-hot probability distribution) not a good characteristic for filtering out low-confidence samples.

To overcome these limitations, we propose an occlusion-robust orientation estimation network by: 1) using a transformer-based network with extensive prior knowledge for joint detection. 2) 23-joint-based human body representation is used to provide additional orientation cues. Thus, our network can recognize and utilize human's visible joints to estimate the orientation. Besides, our network is able to predict reasonable confidence, which is learned by constructing an adversarial strategy between the ground truth and the predicted orientation.

In summary, we propose an orientation estimation network for partial observation scenarios with confidence-aware capability. Through extensive experiments, including two public datasets and a custom-built dataset, we compare to the baseline method [8] and yield state-of-the-art (SOTA) orientation estimation performance. Finally, by integrating our model into an RPF system, we demonstrate our model's superiority in real RPF tasks.

II. RELATED WORK

A. Human Joint Detection

Evidence has demonstrated human joint information is helpful to the HOE task [6] [7] [8]. Since orientation estimation is a top-down process (estimate orientation from a cropped person image), we mainly focus on top-down human joint detection algorithms. HRNet-based algorithms [9] [10] [11] maintain both high-resolution and deep features in multi-stages for human joint detection and achieve the SOTA accuracy on the COCO [12] dataset for about two years. MEBOW adopts the HRNet backbone with human joint detection as the auxiliary task for orientation estimation. Recently, vision transformers have shown great potential in different vision tasks, including human joint detection. Vitpose++ [13] adopts valina vision transformers with a transposed convolution decoder to train the human joint detection model on multiple large-scale datasets with a total of 770k samples. It surpasses the HRNet-based methods by a large margin and shows great generalization ability.

Previous HOE methods often show poor generalization problems due to the small size and low complexity of training data. Additionally, we find that HOE accuracy is highly related to the ability of human joint detection. Therefore, instead of relying on a size-limited orientation dataset, we resort to improving the human joint prediction ability of the network to improve its robustness on HOE even under partial occlusion. Specifically, we utilize a strong backbone [14] that was pre-trained on multiple large-scale joint datasets [12] [15] [16] [17]. Besides, to make the orientation estimation more robust when only the lower body is in the field of view, which is the most common scenario of partial observation, we adopt a human joint representation that contains human foot joints to provide more orientation cues.

B. Deep Learning-based Orientation Estimation

Deep learning is popularly used to solve the problem of orientation estimation. Most orientation datasets are recorded indoors (due to the constraint of motion capture systems) with a simple human movement, a clean background, and full-body or upper-body observations [18] [19] [20]. However, these methods are hard to generalize in the wild applications due to a lack of training samples covering observations of varied environments where partial observation is often observed.

Some research tried to use in-the-wild RGB images to train the HOE task with the help of human joint detection. Yu et al. [6] utilized deep networks to detect human joints and then defined geometric features based on leg, shoulder, and hip joints. The geometric feature is then fed into an SVM model for orientation estimation. Monoloc++ [7] utilized a human joint detection model [21] to detect joints from in-the-wild images, and the joints' positions are then encoded to features to estimate the orientation. Due to the robustness of human joint detection, such a method improved the accuracy by a large margin. However, relying solely on joint information is not enough to provide accurate orientation estimation, particularly in cases where only a partial view of the body is available.

To further improve the accuracy of orientation estimation, MEBOW [8] created a 72-class orientation dataset with 130k in-the-wild RGB samples. With large in-the-wild samples, an end-to-end HRNet [9] with ResNet [22] unit architecture is adopted to train human joint detection and orientation estimation simultaneously. The orientation output is represented as a 72-class one-hot probability vector, resulting in a 5-degree resolution. Due to the similarity between adjacent orientations, MEBOW regresses the orientation probability to a fixed circular Gaussian distribution to make the network converge. With the help of MEBOW's large in-the-wild dataset and the new orientation regression strategy, the MEBOW baseline improves a lot in terms of HOE accuracy under full-body observation.

However, MEBOW's HOE accuracy decreases a lot under partial observation due to limited occlusion data, and the regressions strategy makes the output orientation probability hard to discriminate low-confidence samples. We solve these problems by harnessing the extensive prior knowledge from the human joint detection model. Additionally, inspired by out-of-distribution (OOD) research [23], we propose to learn reliable confidence prediction from training data without explicit confidence labels.

C. Motion-based Orientation Estimation in RPF Systems

For intelligent RPF applications such as walking-aid robots [2] [24] [25] and autonomous shopping carts [4], HOE is necessary since robots need to maintain a desired relative pose to the target person rather than merely keeping a fixed distance. Despite the use of various sensors across different applications, typical RPF systems incorporate target state estimation and robot motion control, as shown in Fig. 1. With different sensors, the robot can detect and track a person

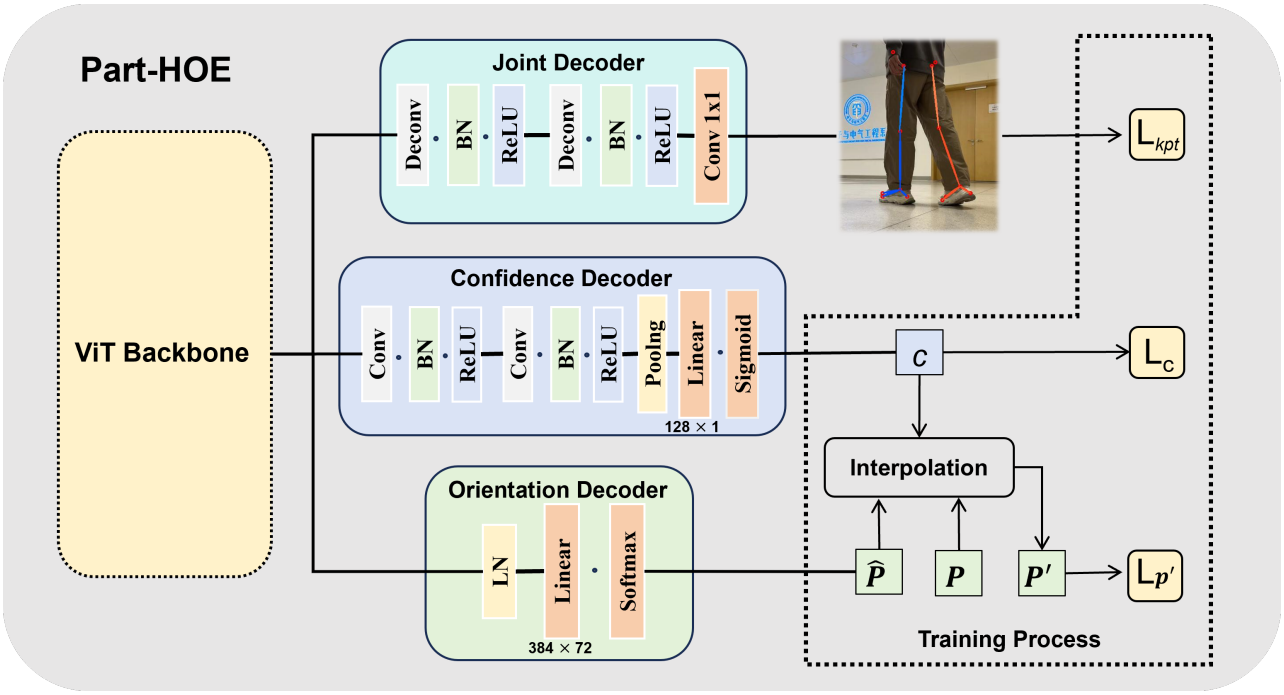


Fig. 2. Our Part-HOE method takes an RGB image as input and extracts features through the ViT backbone. Then, three decoder modules output orientation estimation and confidence estimation, along with 2D human joint detection. Finally, the network is learned by multi-task training.

on the ground plane and further get the state estimation of the target person.

P. Nikdel et al. [4] proposed an autonomous cart that maintains to be in front of a moving human, and the autonomous cart tracks the human's position in the ground plane with a constant motion model. The human orientation is estimated by calculating the direction of the velocity, maintaining observation of (x, y, vt, θ), where orientation $\theta = \arctan\left(\frac{y_t - y_{t-1}}{x_t - x_{t-1}}\right)$. However, such a method suffers when the target person is static or turns with a small position change. Besides, it needs global position information which is obtained from an additional localization algorithm. Qingyang et al. [25] tried to solve the first problem by defining three different motion modes. For example, the orientation of a static human is the perpendicular vector between the left leg and the right leg. However, the state of the human lower limb can be very different for each individual, for example, people can stand by crossing legs. This method is hard to fit the general scenario. Instead of relying on a global frame or leg-specific information, we estimate the human's orientation directly from an image. Our estimation is confidence-aware and reliable even under partial observation. We show that when the RPF system is equipped with our Part-HOE, the following behavior is more consistent than the traditional RPF system in situations of backward and forward following.

III. METHODOLOGY

A. Overview and Problem Statement

To estimate a person's orientation under partial occlusion, we propose a novel orientation estimation method (**Part-HOE**) considering both accuracy and confidence, as shown in Fig. 2. Given a standardized RGB image of a human, we utilize ViT backbone [14] [13] to extract features for the reason that the ViT is pre-trained on multiple large human joint detection datasets and has extensive human joint cues for orientation estimation. The extracted features are then fed into the joint decoder, orientation decoder, and confidence decoder.

The joint decoder uses a simple transpose convolution for up-sampling, followed by a 1×1 convolution layer that predicts the 23 joint heatmaps, including the foot joints (Sec. III-B.1). The orientation decoder is an extremely simple network, which (Sec. III-B.2) is composed of only a normalization layer and a fully connected layer connected to a softmax operation. When joint detection is accurate enough, we found that increasing the complexity of the orientation decoder shows a small benefit to the accuracy improvement. The last decoder is the confidence decoder. By extracting features from the ViT output using convolution, with linear layer and sigmoid operation, the confidence output ranges from 0 to 1. Since there is no explicit confidence label, the confidence estimation is learned in a self-supervised manner as described in Sec. III-B.3. Finally, we integrate our proposed HOE method into an RPF system (Sec. III-C).

B. Part-HOE Model

1) Auxiliary Human Joint Detection: We do not directly use the output of joint detection for orientation estimation; instead, we use the feature map as the input to the orientation decoder to avoid losing other information. The joint detection here is trained as an auxiliary task. Here, we make an effort to enhance the HOE ability by two tricks: 1) applying a ViT-Small backbone with extensive prior knowledge of human

joint detection and 2) adding more human joint constraints for the model to get additional cues from human joints. Transformer-based algorithm ViTPose++ [13] achieved the state-of-the-art (SOTA) human joint detection performance on the COCO [12] dataset. By pre-training on ImageNet with MAE task and conducting human joint detection training on multiple datasets, ViTPose++ shows great generalization ability and accuracy on different datasets. To harness the prior knowledge in the human joint detection area, we use the vision transformer backbone, which is pre-trained on the joint dataset with 770K samples [12] [15] [16] [17].

Baseline MEBOW [8] shows SOTA HOE accuracy under full-body observation. However, the accuracy decreases a lot under partial observation, especially when only lower bodies are observed. To improve the robustness and accuracy under partial observation, especially lower-body observation, we add the additional six-foot joints to the original 17-joint human representation in the COCO [12] dataset to our 23-joint-based human representation.

The loss function for supervising human joint heatmap can be defined as the mean squared error (MSE) between the predicted heat map $\hat{H}$ and the ground truth heat map H. For N human joints and an image with $W \times H$ pixels, the loss function is given by:

$$\mathcal{L}_{kpt} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{W \times H}\sum_{x=1}^{W}\sum_{y=1}^{H}\left(\hat{H}_i(x,y) - H_i(x,y)\right)^2 \tag{1}$$

where $\hat{H}_i(x,y)$ is the predicted intensity of the heatmap at pixel location (x,y) for human joint i, and $H_i(x,y)$ is the corresponding ground truth intensity.

2) Orientation Estimation: Given ViT feature map, the orientation decoder estimates the orientation as a discrete formula same as the baseline [8], denoted as $\hat{\mathbf{p}} = [\hat{p_0}, \hat{p_1}, \ldots, \hat{p_{71}}]$ $(\sum_{i=0}^{71}\hat{p_i} = 1)$, where the max $\hat{p_i}$ indicates the orientation of a person is within the range of $[i \cdot 5^\circ - 2.5^\circ, i \cdot 5^\circ + 2.5^\circ]$. Here, the orientation in a range of $[0^\circ, 360^\circ)$ follows the same definition as MEBOW [8].

We found that when joint detection is accurate enough, increasing the complexity of the orientation decoder yields no benefit. Therefore, our orientation decoder is composed of only a normalization layer with a fully connected layer connected with softmax operation as shown in the orientation decoder in Fig. 2.

We converted the orientation labels $l_{\text{gt}} \in [0, 71] \cap \mathbb{Z}$ to "circular" Gaussian probability $\mathbf{p}$ as MEBOW [8], $\mathbf{p} = [p_0, p_1, \ldots, p_{71}](\sum_{i=0}^{71} p_i = 1)$ and trained it as a regression task:

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

where σ is a constant value, and the "circular" Gaussian probability is visualized in Fig. 3.

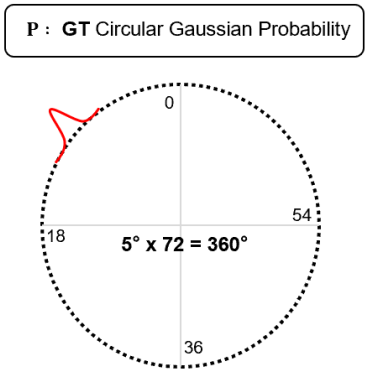


Fig. 3. An explanation of Circular Gaussian Probability in the interpolation operation of Part-HOE.

3) Confidence Estimation: The confidence decoder outputs a confidence value range from zero to one, indicating if the orientation estimation is reliable. As we mentioned in Sec. III-B.2, the training of orientation is a regression task since the loss could not converge if using traditional classification loss, i.e. cross-entropy loss. Therefore, the model regresses orientation probability to constant Gaussian distribution (Fig. 3), which means the output $\hat{\mathbf{p}}$ is a regressed distribution instead of a probability. In partial observation scenarios, the model tends to predict orientations with the same distribution even when the prediction is highly unreliable. Inspired by [23], we proposed a confidence-aware orientation estimation method to predict the confidence of each orientation estimation as shown in Fig. 2. Even without the annotation of confidence that is available during training, confidence can be learned by constructing an adversarial strategy between the ground truth and the predicted orientation. Denote the predicted confidence $c \in (0, 1)$, and construct $\mathbf{p}'$ for training confidence c and orientation $\mathbf{p}$ at the same time:

$$\mathbf{p}' = c \cdot \hat{\mathbf{p}} + (1 - c) \cdot \mathbf{p} \tag{3}$$

The loss function of the constructed $\mathbf{p}'$ is defined as:

$$\mathcal{L}_{\mathbf{p}'} = \sum_{i=0}^{71}(p_i' - p_i)^2 \tag{4}$$

A penalty $\mathcal{L}_c$ for confidence c along with $L_{\mathbf{p}'}$ for preventing confidence c approaching 0. The penalty loss is expressed as:

$$\mathcal{L}_c = -log(c) \tag{5}$$

If the observation is reliable, the confidence c will converge towards one; conversely, confidence c will approach zero.

The final loss function is the sum of the orientation loss, the penalty confidence loss, and the joint detection loss. Here, λ is a weight coefficient that dynamically changes with $\mathcal{L}_c$, used to balance the loss value between orientation and confidence.

$$\mathcal{L} = \mathcal{L}_{\mathbf{p}'} - \lambda \cdot \mathcal{L}_c + \mathcal{L}_{kpt} \tag{6}$$

TABLE I. Evaluation of orientation accuracy on three datasets. Acc (N°) is the estimation accuracy that estimation is regarded as true if the estimated orientation error is within (-N°, +N°). MAE (°) is the mean absolute error of orientation. The data inside the bracket is evaluated under full observation, while those outside the bracket are evaluated under partial observation.

Dataset	Methods	GFlops ↓	Params ↓	Acc (5°) ↑	Acc (15°) ↑	Acc (30°) ↑	MAE (°) ↓
MEBOW	MEBOW	8.28G	39.6M	47.0% (68.6%)	76.6% (90.7%)	90.0% (96.9%)	16.3 (8.4)
	Monoloco++	8.70G	25.5M	42.6% (49.8%)	75.0% (81.5%)	91.3% (93.5%)	14.6 (12.5)
	Ours	**5.62G**	**24.2M**	**52.4%** (**70.1%**)	**82.0%** (**91.2%**)	**93.4%** (**96.9%**)	**13.4** (**8.1**)
H3.6M	MEBOW	8.28G	39.6M	11.7% (30.8%)	31.7% (**77.8%**)	55.6% (98.3%)	43.8 (10.0)
	Ours	**5.62G**	**24.2M**	**18.7%** (**32.4%**)	**49.8%** (77.1%)	**77.8%** (98.3%)	**21.7** (**9.9**)
Custom-built	MEBOW	8.28G	39.6M	12.0%	37.1%	66.8%	34.7
	Ours	**5.62G**	**24.2M**	**17.7%**	**47.8%**	**83.4%**	**21.0**

C. Robot Person Following System

We implement an RPF system that can follow the target person both forward and backward to demonstrate the importance of Part-HOE in real RPF applications. Following forward or backward requires the robot to maintain a fixed distance in the front or back of the target person. The location of the target person is estimated using a leg tracker [5]. Denote the pose of the target person as $(x_{ped}, y_{ped}, \theta_{ped})$ in the robot's local frame. The ideal following backward pose $(x_{backward}, y_{backward}, \theta_{backward})$ is calculated by:

$$\begin{aligned} x_{backward} &= x_{ped} - \cos(\theta_{ped}) \\ y_{backward} &= y_{ped} - \sin(\theta_{ped}) \\ \theta_{backward} &= \theta_{ped} \end{aligned} \tag{7}$$

The ideal following forward pose is calculated by:

$$\begin{aligned} x_{forward} &= x_{ped} + \cos(\theta_{ped}) \\ y_{forward} &= y_{ped} + \sin(\theta_{ped}) \\ \theta_{forward} &= \theta_{ped} + \pi \end{aligned} \tag{8}$$

Both following forward and backward tasks need to estimate the position and orientation of the target person accurately. The RPF system includes a state estimation module and robot control module as shown in Fig. 1. The details of other modules can be found in our previous work [26] [27]. Here, we integrated our Part-HOE network into the state estimation module of the RPF system.

IV. EXPERIMENTS

To validate the accuracy and robustness of our proposed Part-HOE model, we conducted comparisons on three different orientation datasets, and we also evaluated it in RPF tasks by integrating our method into a robot system. In this section, we first introduce the datasets, the baselines, and the details of the implementation of our experiments. Secondly, we demonstrate the superiority of our Part-HOE in terms of HOE accuracy and discrimination ability for low-confidence samples. An ablation study is conducted to verify the importance of different modules. Lastly, we show that Part-HOE can improve the robustness and consistency of RPF tasks.

A. Datasets

In this work, three datasets are used in the experiment, including two public datasets (MEBOW [8] dataset and Human-3.6M [28]) and our custom-built dataset.

- The MEBOW dataset contains 130k in-the-wild samples, and the orientation is annotated to 72 class labels. The label i indicates the orientation of the person is within the range of $[i \cdot 5° - 2.5°, i \cdot 5° + 2.5°]$, resulting in a 5-degree resolution.
- The Human3.6 M dataset provides 3D human joint annotations for continuous monocular image sequences containing different actions. To validate our model in the continuous image sequence, we also evaluated our model on the walking sequences with 158K full-body observation samples in total, where the orientations range from $[0°, 360°)$ can be calculated by following the definition of MEBOW [8].
- Our custom-built dataset includes 5k partial observation images recorded under a motion capture system with a helmet and a robot equipped with an RGB-D camera and motion capture markers. The orientation annotation is continuously calculated by the transformation between the helmet and the robot, and it is recorded while the robot performs a real RPF task.

B. Baseline Methods

For comparison in terms of orientation estimation accuracy, we compared with the strong baseline MEBOW and Monoloco++. We also conducted multiple comparisons of confidence and improvement in RPF tasks with the current SOTA MEBOW.

- MEBOW [8]: We compare the accuracy of the MEBOW baseline method on its validation dataset and the Human3.6M Walking dataset. To get the partial observations, we cropped only the image lower than the hip joint for evaluation.
- Monoloco++ [7]: The Monoloco models take 2d human joint as input and output continuous orientation in a range of $[-\pi, \pi]$. To make a fair comparison, we retrain the monoloco network on the MEBOW dataset and provide input with ground truth human joint.

We also make a comparison with the traditional RPF algorithm [25] [4] [5], where the state estimation of humans

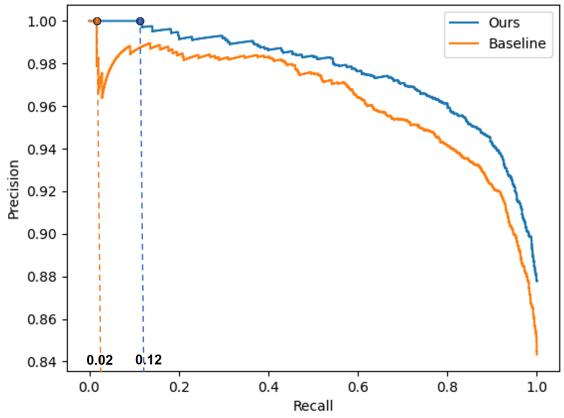


Fig. 4. Precision-recall curve under partial observation. The dashed lines indicate the max recall at 100% precision.

is tracked using the constant velocity model. To exclude the influence of position estimation, we directly obtain the ground truth of the human's positions from a motion capture system in 10 Hz.

C. Implementation Details

The implementation of our model is based on ViTpose++ [13] and MEBOW [8], and the backbone was pre-trained on the human joint detection task. ViT-Small is used as our backbone, considering the computation cost. For the MEBOW dataset and Human3.6M, inputs were processed by cropping the target person using the ground truth bounding box, followed by a standardized operation. Our custom-built dataset employed YOLOX [29] for bounding box detection, subsequently applying the same standardization procedure. Our model is trained on the MEBOW dataset [8], where augmentation techniques, including cropping, flipping, and scaling, are used during the training process. We use Adamw optimizer (learning rate is equal to 0.001) to train our Part-HOE for 80 epochs.

For RPF implementation, a Clearpath Dingo-O and a laptop with Intel(R) Core(TM) i5-10200H CPU @ 2.40GHz and NVIDIA GeForce GTX 1650 are used. A Realsense D435i with 1280 $\times$ 720 resolution and 30 Hz frequency is mounted on the robot with a 30° tilt angle and a height of 0.17m relative to the ground plane. To exclude the influence of other modules, we utilize a motion capture system to obtain the human's positions, and the orientation is estimated by the baseline MEBOW or our Part-HOE. For the traditional RPF implementation, we follow the code in [5]. The control algorithm employs model predictive control, executing the robot person following both backward and forward. (see Sec. III-C)

D. Experimental Results

1) Evaluation of Orientation Accuracy: The evaluation of orientation accuracy is conducted on three datasets: the MEBOW dataset, the Human 3.6M Dataset, and our custom-built dataset. Two metrics are used to evaluate the orientation accuracy, including the percentage of error within n°

TABLE II. Ablation study for the effectiveness of additional foot joint constraints and pre-trained ViT backbone

method	ViT	foot joints	Acc (30°) ↑	MAE (°) ↓
MEBOW	✗	✗	90.0%	16.3
	✗	✓	90.3% (+0.3%)	16.2 (-0.1)
	✓	✗	92.4% (+2.4%)	14.0 (-2.3)
Ours	✓	✓	93.4% (+3.4%)	13.4 (-2.9)

and the mean absolute error (MAE). We also evaluate the model computation cost, considering it is used on real robot tasks. In terms of computational cost, our model achieves 32% fewer Flops and 39% fewer parameters compared to the baseline MEBOW as shown in Table I. In terms of orientation estimation accuracy, we evaluate our method on both full-body and partial-body observation, and our method shows better orientation accuracy in both scenarios. Notably, in scenarios of partial observations, our model achieves +4%, +22%, and +16% improvement on three orientation datasets as shown in Table I.

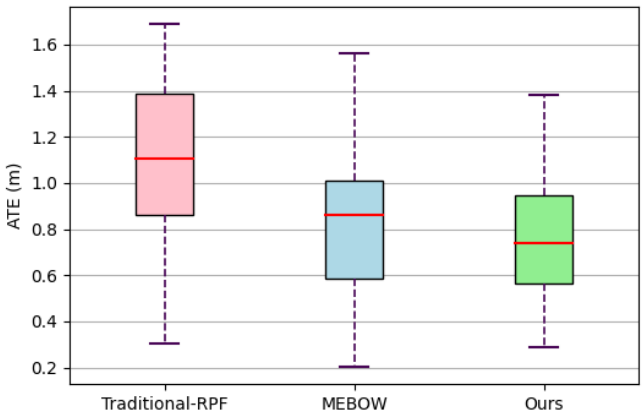
2) Ablation Study: For the accuracy contribution to the whole model, we conducted ablation experiments for the ViT backbone and foot joints. The results in Table II show that ViT backbone and foot joint constraints contribute to +2.4% and +0.3% improvement, respectively, under the partial observation of the MEBOW dataset.

3) Evaluation of Confidence: The confidence in the range of $(0, 1)$ indicates the reliability of orientation estimation, and the baseline orientation output $\hat{\mathbf{p}}$ represents a different probability of 72 class orientation. To evaluate the predicted confidence with the baseline method, we can use the orientation probability and confidence to classify the reliable and unreliable samples. Here, we utilize max recall @ 100% precision as evaluation metrics for binary reliability classification accuracy. The max recall @ 100% precision indicates the ability to find true orientation estimation without false positives. This ability is crucial for RPF because trusting a wrong estimation would result in dangerous RPF behavior.

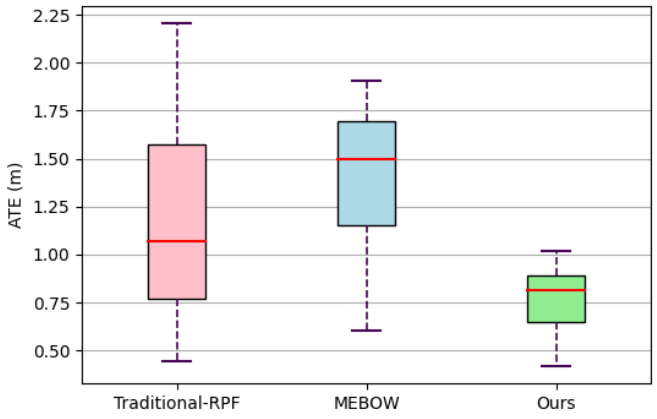
Here, we set the samples with an orientation estimation error bigger than 20° as unreliable samples. We conduct experiments under a partial observation situation in the MEBOW dataset. The baseline MEBOW dropped from the very beginning 0.02 while we are able to keep 100% precision till recall is 0.12 as shown in Fig. 4,

The baseline method makes wrong predictions with an unexpectedly high probability, and we can observe from Fig. 6 that MEBOW tends to give high probability output when there are enough visible joints and ignores the fact that only a few orientation cues can be observed (we scale the probability relative to its maximum value). Although such a scenario is hard for HOE because most orientation cues are occluded, our method can give reasonable confidence output and predict accurate orientation estimation.

4) Real Robot Experiments: To further evaluate our model's accuracy and robustness, we integrate part-HOE into a robot person following (RPF) system as described in



(a) Backward RPF



(b) Forward RPF

Fig. 5. Comparison of different RPF methods in real robot experiments. This comparison shows the absolute trajectory error (ATE) for different RPF methods, with the green box representing our method, the blue box representing MEBOW, and the red box representing the traditional RPF method that relies solely on human velocity for orientation. The lines above and below the dashed lines indicate the maximum and minimum values, while the red line represents the mean. In the two person-following scenarios, which include (a) Backward RPF task evaluation and (b) Forward RPF task evaluation, using PartHOE for orientation estimation demonstrates the best performance in person-following.

Sec. III-C. There are two metrics used in real robot experiments: the absolute trajectory error (ATE) of the following trajectory and the orientation estimation accuracy after using confidence. For the RPF task evaluation, we define the following forward and backward tasks, which correspond to the two cases where the robot accompanies the target person forward and backward at a fixed distance. Here, we set the following distance to 1 m. By integrating our Part-HOE into the RPF system, we significantly reduce ATE by 0.65 m at the following forward task and reduce ATE by 0.31 m at the following backward task compared to the traditional RPF system, as shown in Fig. 5 (b). We also compare MEBOW with the same experiment setting, and MEBOW achieves 1.23 m ATE at the following forward task and 0.82 m ATE at the following backward task. The MEBOW-based RPF system is better than the traditional RPF system but worse than our method. We can see that our following trajectory is closer to the ground truth compared to MEBOW, as shown in Fig. 5 (a).

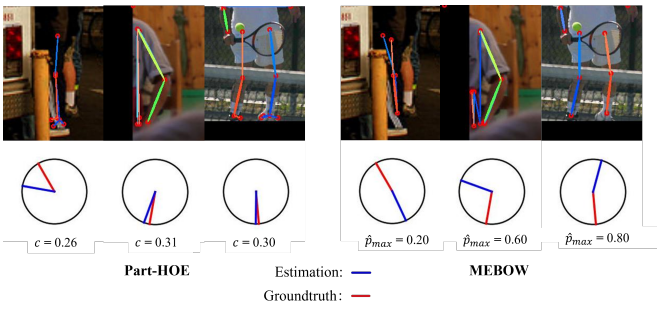


Fig. 6. MEBOW tends to output a high probability but incorrect prediction, as shown in the example above, while our Part-HOE provides a more reasonable confidence level. The variable c represents the confidence output of PartHOE, while $\hat{p}_{\max}$ denotes the maximum probability output of MEBOW.

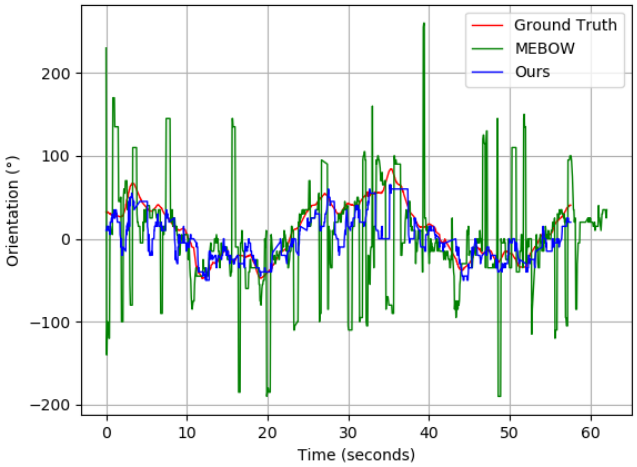


Fig. 7. Visualization of HOE during forward RPF task.

We further evaluate the confidence output of our orientation estimation model. The unreliable samples are filtered out during the following process, and we just directly set the orientation at the current frame to the most confident estimation among the last 5 frames since the frequency of our model is about 25 FPS. The orientation filtered by our confidence achieves higher accuracy compared to the orientation filtered by MEBOW probability and shows slight variation as shown in Fig. 7.

V. CONCLUSION

In this paper, we show that enhanced joint detection, particularly with a pre-trained ViT model and additional foot joints, significantly improves orientation accuracy under occlusion, with gains of up to 22% on the Human3.6M dataset and 16% on our custom dataset. The proposed self-supervised method for confidence estimation offers more reliable filtering of uncertain samples compared to traditional approaches like MEBOW. Additionally, the proposed Part-HOE method demonstrates superiority in real robot applications, i.e., robot person following. However, Part-HOE has yet to fully utilize temporal information and still struggles with effectively filtering out unreliable orientation estimates. We aim to address these issues in our future work.

REFERENCES

[1] M. J. Islam, J. Hong, and J. Sattar, "Person-following by autonomous robots: A categorical overview," *The International Journal of Robotics Research*, vol. 38, no. 14, pp. 1581–1618, 2019.

[2] K. Wakita, J. Huang, P. Di, K. Sekiyama, and T. Fukuda, "Human-walking-intention-based motion control of an omnidirectional-type cane robot," *IEEE/ASME Transactions on Mechatronics*, vol. 18, no. 1, pp. 285–296, 2013.

[3] A. Ashtari, S. Stevšić, T. Nägeli, J.-C. Bazin, and O. Hilliges, "Capturing subjective first-person view shots with drones for automated cinematography," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 5, pp. 1–14, 2020.

[4] P. Nikdel, R. Shrestha, and R. Vaughan, "The hands-free push-cart: Autonomous following in front by predicting user trajectory around obstacles," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 4548–4554.

[5] A. Leigh, J. Pineau, N. Olmedo, and H. Zhang, "Person tracking and following with 2d laser scanners," in *2015 IEEE International Conference on robotics and automation (ICRA)*. IEEE, 2015, pp. 726–733.

[6] D. Yu, H. Xiong, Q. Xu, J. Wang, and K. Li, "Continuous pedestrian orientation estimation using human keypoints," in *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2019, pp. 1–5.

[7] L. Bertoni, S. Kreiss, and A. Alahi, "Perceiving humans: from monocular 3d localization to social distancing," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 7401–7418, 2021.

[8] C. Wu, Y. Chen, J. Luo, C.-C. Su, A. Dawane, B. Hanzra, Z. Deng, B. Liu, J. Z. Wang, and C.-h. Kuo, "Mebow: Monocular estimation of body orientation in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[9] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5693–5703.

[10] H. Liu, F. Liu, X. Fan, and D. Huang, "Polarized self-attention: Towards high-quality pixel-wise regression," *arXiv preprint arXiv:2107.00782*, 2021.

[11] C. Wang, Y. Zhou, F. Zhang, and P. Mok, "Unbiased feature position alignment for human pose estimation," *Neurocomputing*, vol. 537, pp. 152–163, 2023.

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

[13] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "Vitpose++: Vision transformer for generic body pose estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[15] M. Andriluka, L. Pishchulin, P. Gehler, and B. 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*, 2014, pp. 3686–3693.

[16] J. Wu, H. Zheng, B. Zhao, Y. Li, B. Yan, R. Liang, W. Wang, S. Zhou, G. Lin, Y. Fu, *et al.*, "Ai challenger: A large-scale dataset for going deeper in image understanding," *arXiv preprint arXiv:1711.06475*, 2017.

[17] S. Jin, L. Xu, J. Xu, C. Wang, W. Liu, C. Qian, W. Ouyang, and P. Luo, "Whole-body human pose estimation in the wild," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. Springer, 2020, pp. 196–214.

[18] M. Andriluka, S. Roth, and B. Schiele, "Monocular 3d pose estimation and tracking by detection," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 623–630.

[19] B. Lewandowski, D. Seichter, T. Wengefeld, L. Pfennig, H. Drumm, and H.-M. Gross, "Deep orientation: Fast and robust upper body orientation estimation for mobile robotic applications," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 441–448.

[20] T. Fischer, H. J. Chang, and Y. Demiris, "Rt-gene: Real-time eye gaze estimation in natural environments," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 334–352.

[21] S. Kreiss, L. Bertoni, and A. Alahi, "Openpifpaf: Composite fields for semantic keypoint detection and spatio-temporal association," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 13 498–13 511, 2021.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[23] T. DeVries and G. W. Taylor, "Learning confidence for out-of-distribution detection in neural networks," *arXiv preprint arXiv:1802.04865*, 2018.

[24] J. Huang, P. Di, T. Fukuda, and T. Matsuno, "Motion control of omni-directional type cane robot based on human intention," in *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2008, pp. 273–278.

[25] Q. Yan, J. Huang, Z. Yang, Y. Hasegawa, and T. Fukuda, "Human-following control of cane-type walking-aid robot within fixed relative posture," *IEEE/ASME Transactions on Mechatronics*, vol. 27, no. 1, pp. 537–548, 2022.

[26] H. Ye, J. Zhao, Y. Pan, W. Chen, L. He, and H. Zhang, "Robot person following under partial occlusion," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 7591–7597.

[27] H. Ye, J. Zhao, Y. Zhan, W. Chen, L. He, and H. Zhang, "Person re-identification for robot person following with online continual learning," *IEEE Robotics and Automation Letters*, pp. 1–8, 2024.

[28] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.

[29] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.