AiOS: All-in-One-Stage Expressive Human Pose and Shape Estimation

Supplementary Material

1. Overview

Due to the limited space on the main paper, we present more additional details in this supplementary material, covering the following aspects:

- Additional details on datasets, experimental parameters related to data augmentation, and implementation details in Sec. 2.
- Supplementary experiments investigating the impact of bounding boxes on algorithm performance in Sec. 3.
- Extra SOTA quantitative and qualitative comparison experiments on EgoBody-EgoSet [22], BEDLAM [1], and ARCTIC [6], and extra visualization comparison on single person, multiperson, and synthetic images. in Sec. 4.

2. Experiment Setup

2.1. Datasets

This subsection primarily describes the characteristics of the datasets utilized in our main paper and provides details on how we use these datasets in the training and testing stages.

AGORA [16] is a synthetic image dataset that encompasses lots of complex scenarios, including severe occlusion and truncations. It has 14K training images with 122K instances and 1K validation images with around 8K instances. Recently, AGORA has become an essential benchmark for tasks related to SMPL [12] and SMPL-X [17], primarily for its effectiveness in assessing algorithm performance in occlusion-heavy scenes. Our approach utilizes the complete dataset for training, validation, and testing purposes.

BEDLAM [1] is a synthetic video dataset offering a wide variety of data, including diverse body shapes, motions, skin tones, hairstyles, and clothing. The clothing is notably rendered using a professional physics simulator, enhancing realism, particularly in depicting character movement. Each image in the dataset features between 1 to 10 individuals. Originally comprising 286K images, we downscaled it by a factor of 5, yielding 57L images with 190K instances, which we then utilized for training.

MSCOCO [10] is a large-scale real-world dataset designed for object detection, segmentation, keypoint detection, and captioning. On the basis of the keypoint subset, we use the pseudo SMPL-X [17] annotations from NeuralAnnot [14] for training. It contains 56K multi-person images, featuring a total of 147K instances.

EHF [17] is an indoor dataset consisting of only 100 images along with corresponding SMPL-X labels. Due to the absence of a corresponding training set, current algorithms often utilize it to assess algorithm generalization. It only contains test sets with 100 single-person images.

UBody [8] is a dataset containing diverse real-life scenarios, including movies, TV shows, talk shows, vlogs, sign language, online classes, and more. UBody contains an extensive collection of rich gestures and expressions that are not present in other real-world datasets. We down-sample the training set to 54K images with 66K instances. We utilize a downsampled test set, as used in SMPLer-X[2] and OSX [8], which has 2642 images with 2642 instances.

ARCTIC [6] is an indoor dataset which mainly focuses on the hand-object interaction. We downsample the original training set 1000 times for a train set of 50K images with 50K instances. We discard the egocentric view, which only contains hands, in our training. For the test set, following [2], we keep the original test set with 207K images to evaluate our method.

Egobody [22] is a large-scale dataset for egocentric views. The egocentric view datasets are collected using Microsoft HoloLens 2, including RGB, depth, eye gaze, head tracking, and hand tracking. The SMPL-X data is obtained by fitting the multi-Kinect rig data. We downsample 2 times to get 45K images with 45K instances in the egocentric-view split for training. The test set has 62K images.

We pre-train our model by randomly sampling data from AGORA, BEDLAM, and COCO. We choose these three datasets because their multi-person images satisfy our method of perceiving the positions of individuals in the images. Additionally, AGORA and BEDLAM have the advantage of accurate ground truth, while COCO provides the diversity of real-world scenes, preventing our method from overfitting to synthetic data.

SMPLer-X [2] indicates that scaling up the data can enhance algorithm performance. Hence, we incorporate Egobody, ARCTIC, and UBody data into our training. The sampling probabilities are 0.2, 0.2, 0.2, 0.2, 0.1, and 0.1, respectively, for AGORA, BEDLAM, COCO, Egobody, ARCTIC, and UBody. ARCTIC and Egobody provide more accurate real-world whole-body data compared to COCO, while UBody contributes abundant and diverse gestures and expressions. All the datasets are standardized into the HumanData [5] format. Some visual demonstrations of AiOS are provided in Fig. 1 and Fig. 2.

2.2. Implementation details

The training is conducted on 16 V100 GPUs, with a total batch size of 32. We initialize AiOS's backbone, encoder, and part of the decoder from a weight trained on COCO human pose estimation, provided by ED-Pose [21]. We use Adam optimizer with a step learning rate for training. We

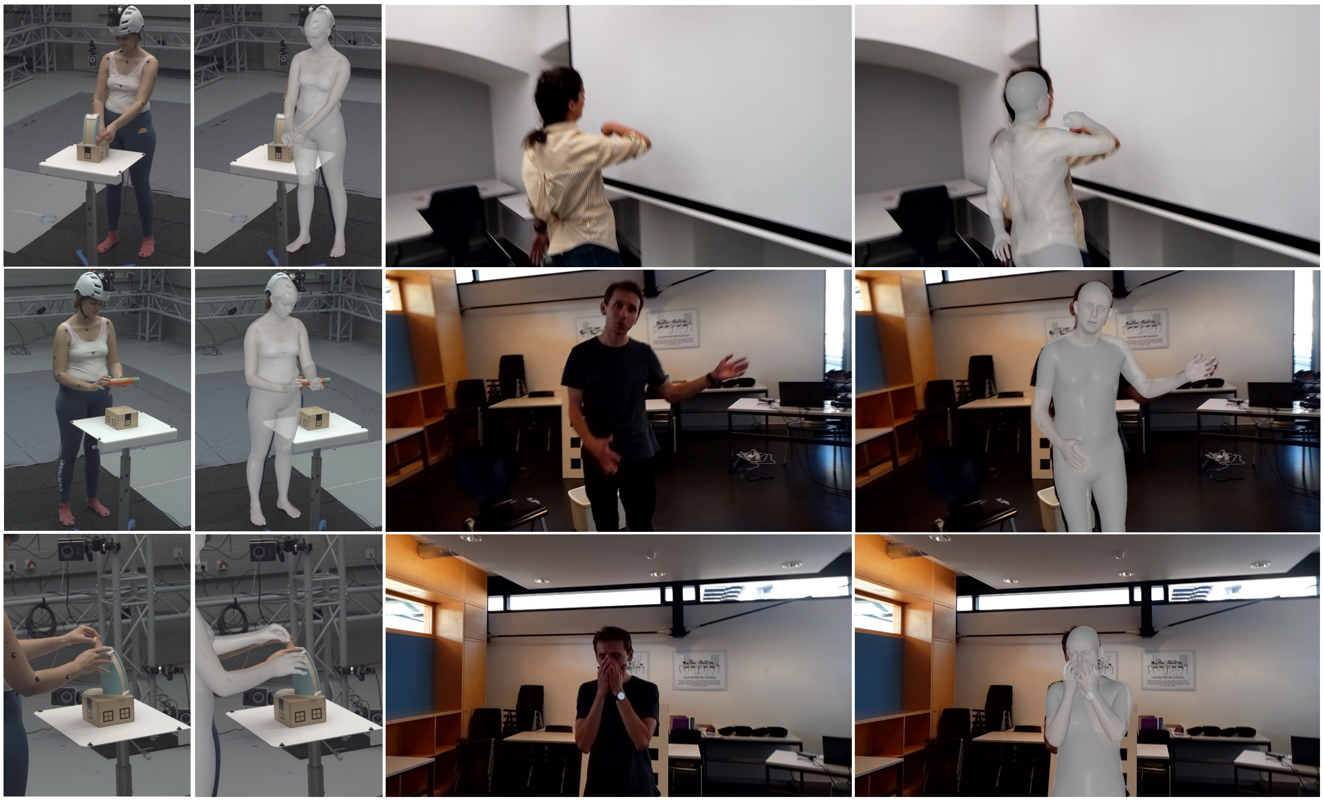

Figure 1. **Illustration of AiOS in indoor datasets.** The first two columns are the qualitative results on ARCTIC [6], while the last two columns are the qualitative results on Egobody [22].

first train our model for 60 epochs with an initial learning rate 1×10^{-4} and drop at the 50th epoch on AGORA, BEDLAM, and COCO. We finetune it for 50 epochs with 1×10^{-5} initial learning rate and drop at the 25th epoch on all train datasets.

During the training stage, we adopt color jittering, random horizontal flipping, random image resizing, and random instance cropping. For the color jittering, we randomly apply a variation of ± 0.2 in the RGB channels. For the random horizontal flipping, we augment images and their corresponding annotations by horizontally flipping them with a probability of 0.5. For random image resizing, during the training process, we resize the images in proportion, ensuring that the shorter side is kept between 480 and 800 pixels and the longer side does not exceed 1333 pixels. During testing, we set the shorter side to 800 pixels, and the longer side is scaled proportionally, with the constraint that it does not exceed 1333 pixels. For random instance cropping, we first randomly enable the cropping operation with a probability of 0.5. When performing the cropping operation on a multi-person dataset, such as AGORA [16], BEDLAM [1], COCO [10], We randomly sample 1 to N instances from the image with probability 0.5, where N is the total number of people in the image. The image is then cropped according to their collective bounding box. Alternatively, with a 0.5 probability, the cropping operation is applied to the collective bounding box of all instances. We maintain the original aspect ratio during the cropping process to avoid cropping unusual aspect ratios.

2.3. Loss Functions

Body-location Decoder. The losses for supervising Body-location Decoders are L_{box} and L_{cls}. L_{box} contains L1 loss and the GIOU loss [18] for the body location. L_{cls} is focal loss [9] for classify body tokens.

Body-refinement Decoder.. The losses for supervising Body-refinement Decoders are L_{box}, L_{j2d} and L_{smplx_b}. L_{box} contains L1 and loss GIOU loss for the body location, hands location, and face location. L_{j2d} is the L1 loss and OKS loss supervising the body joints location and L_{smplx} contains L1 loss with ground truth SMPL-X body parameters L_{param}, L1 loss L_{kp3d} for 3D body keypoints regressed by SMPL-X J-regressor, and L1 loss L_{kp2d} for projected 2D keypoints.

Wholebody-refinement Decoder.. The losses for supervising Wholebody-refinement Decoders are L_{j2d} and L_{smplx}. L_{j2d} is the L1 loss and OKS loss supervising the whole-body joints location and L_{smplx} contains L1 loss with ground truth

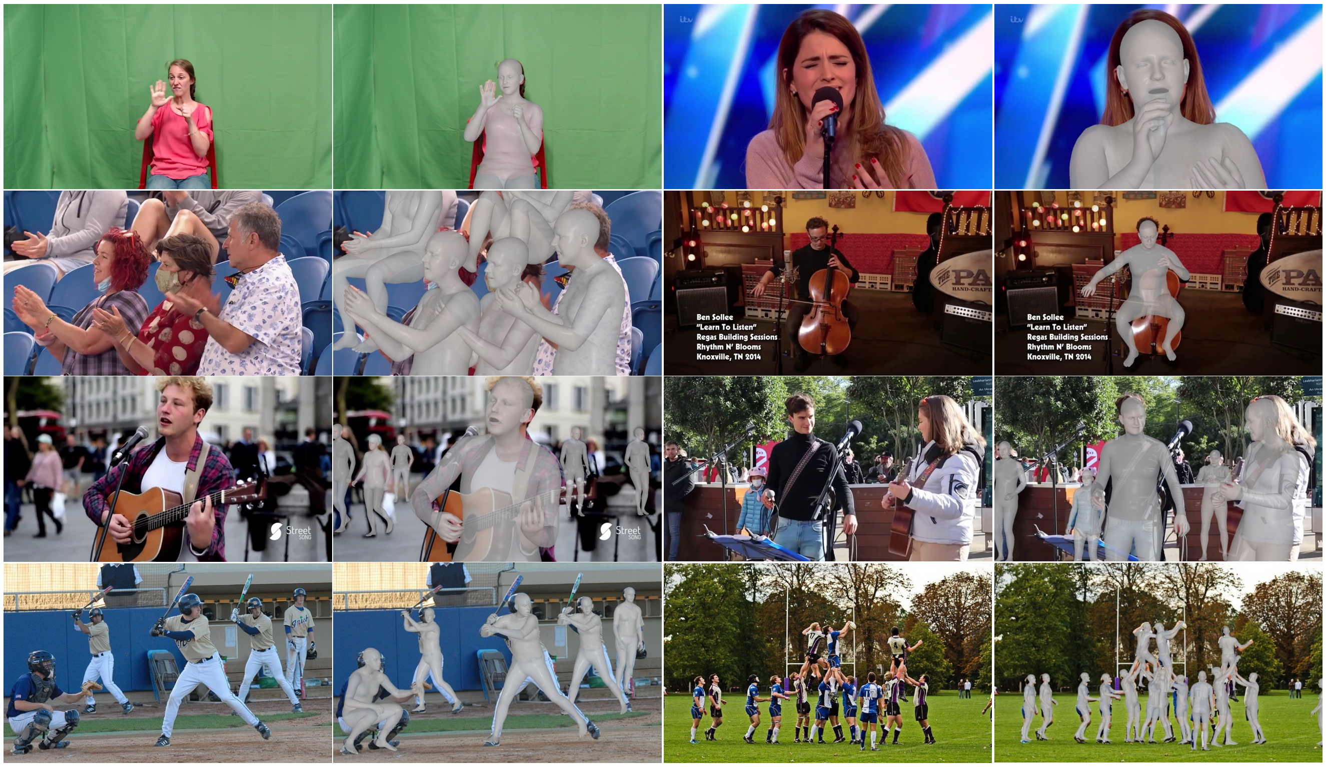

Figure 2. **Illustration of AiOS in outdoor datasets.** The first three rows are qualitative results on UBody [8], and the last row is qualitative results on COCO [10]

SMPL-X parameters L_{param}, L1 loss L_{kp3d} for 3D whole-body keypoints regressed by SMPL-X J-regressor, and L1 loss L_{kp2d} for projected 2D whole-body keypoints.

Loss Weights We weighted-sum all the losses at all stages as the final loss. The loss weights of the same type of loss in different stages are the same, which are shown as follows: L_{cls}: 2.0, L_{smplx}: 1.0 (pose), 0.01 (shape, expression), L_{kp3d}: 1.0 (body), 0.5 (face, hand), L_{kp2d}: 1.0 (body), 0.5 (face, hand), L^{1}_{j2d}: 10, L_{oks}: 4.0 (body), 0.5 (face), 0.5 (hands), L_{giou}: 2.0, L^{1}_{box}: 5.0.

3. Sensitivity Analysis of Performance to Bounding Box Accuracy

In this section, we present that current methodologies exhibit a significant sensitivity to bounding boxes. Our experiments are carried out on the AGORA [16] validation set, utilizing official evaluation tools [1] for a thorough assessment. We report several key metrics, including Normalized Mean Vertex Error (NMVE), Normalized Mean Joint Error (NMJE), Mean Vertex Error (MVE), and Mean Per-Joint Position Error (MPJPE), which focus on the reconstruction accuracy of body, hands, and face. Additionally, we include F-Score, precision, and recall to evaluate the accuracy of detection.

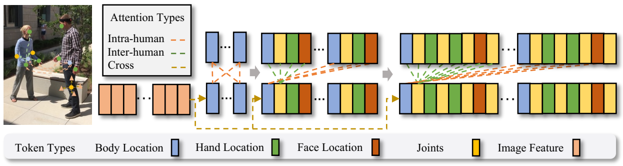

Figure 3. **Illustration of Inter-human, Intra-human, and Cross Attention.** Inter-human attention is conducted between the body tokens and body, hand, and face location tokens of different persons'. Intra-human attention is conducted with the location and joint tokens of the same person. Cross-attention is conducted between image features and all the tokens.

Firstly, we utilize the ground truth (GT) bounding box to crop the image and evaluate the performance of OSX [8] and SMPLer-X [2], denoted by **GT Box** in Table 1. Following this, we employ DAB-DETR[3, 11], an off-the-shelf detection model, to identify human bounding boxes, replacing the GT boxes for image cropping. We present detection results under three bounding box score thresholds: 0.1, 0.2, and 0.3, labeled as detected boxes with scores 0.1, 0.2, and 0.3, respectively. A higher score indicates greater confidence in the detection results, but it may lead to missing instances with severe occlusion. This is reflected in the metrics as high accuracy but low recall. Conversely, a lower score threshold retains results with lower confidence, capturing

[1] https://github.com/pixelite1201/agora_evaluation

Methods	Box Type	F Score↑	Precision↑	Recall↑	NMVE↓ (*mm*)		NMJE↓ (*mm*)		MVE↓ (*mm*)					MPJPE↓ (*mm*)				
					All	Body	All	Body	All	Body	Face	LHand	RHand	All	Body	Face	LHand	RHhand
OSX [8]	GT box	1.0	1.0	1.0	120.4	75.2	116.1	72.1	120.4	75.2	39.4	47.5	48.8	116.1	72.1	40.5	45.1	46.4
	Detected box *w* score 0.1	0.58	0.42	0.92	237.8	155.9	228.3	147.4	137.9	90.4	43.8	48.3	50.4	132.4	85.5	46.4	46.1	48.1
	Detected box *w* score 0.2	0.8	0.74	0.87	165.9	110.6	159.2	104.7	132.7	88.5	40.7	44.9	46.9	127.4	83.8	43.1	42.9	44.8
	Detected box *w* score 0.3	0.86	0.89	0.83	148.1	100.3	142.1	95.1	127.4	86.3	37.8	41.8	43.6	122.2	81.8	39.8	39.8	41.6
	GT box$^{\times}$	0.95	0.93	0.97	123.6	77.3	119.3	74.2	117.4	73.4	38.2	46.3	47.7	113.3	70.5	39.0	43.9	45.3
	AiOS box	0.95	0.93	0.97	124.3	78.7	120.0	75.7	118.1	74.8	37.4	45.5	47.0	114.0	71.9	38.3	43.2	44.7
SMPLer-X [2]	GT box	1.0	1.0	1.0	99.5	60.2	95.5	57.5	99.5	60.2	32.9	41.9	43.0	95.5	57.5	34.1	39.5	40.5
	Detected box *w* score 0.1	0.58	0.43	0.92	203.4	131.6	195.3	124.7	118.0	76.3	37.1	43.6	44.4	113.3	72.3	39.5	41.3	42.2
	Detected box *w* score 0.2	0.81	0.75	0.88	140.7	92.6	135.1	87.8	114.0	75.0	34.7	40.7	41.7	109.4	71.1	36.9	38.5	39.5
	Detected box *w* score 0.3	0.86	0.9	0.83	126.4	84.4	121.3	80.1	108.7	72.6	31.8	37.8	38.7	104.3	68.9	33.7	35.7	36.7
	GT box$^{\times}$	0.95	0.93	0.97	101.5	61.4	97.6	58.8	96.4	58.3	31.7	40.8	41.9	92.7	55.9	32.6	38.3	39.4
	AiOS box	0.95	0.93	0.97	103.3	63.5	99.6	61.1	98.1	60.3	31.3	40.4	41.8	94.6	58.0	32.3	38.0	39.4
AiOS	-	0.95	0.93	0.97	106.4	64.2	103.4	62.1	101.1	61.0	30.7	43.9	45.7	98.2	59.0	32.8	41.5	43.4

Table 1. **AGORA validation set.** OSX [8] and SMPLer-X [2] are finetuned on the AGORA training set. However, AiOS is not intentionally fine-tuned exclusively on AGORA. **GT Box** means that this method uses the ground truth bounding box to crop the image. **GT box**$^{\times}$ means that this method uses the ground truth bounding box to crop the image but filters the instances that AiOS fails to detect. **AiOS box** means that this method uses the bounding box provided by AiOS. The best results are colored with red, and the second-best results are colored with blue for OSX and SMPLer-X, respectively.

Methods	Box Type	F Score↑	Precision↑	Recall↑	NMVE↓ (*mm*)		NMJE↓ (*mm*)		MVE↓ (*mm*)					MPJPE↓ (*mm*)				
					All	Body	All	Body	All	Body	Face	LHand	RHand	All	Body	Face	LHand	RHhand
OSX [8]	GT box$^{\times}$	0.95	0.93	0.97	123.6	77.3	119.3	74.2	117.4	73.4	38.2	46.3	47.7	113.3	70.5	39.0	43.9	45.3
	GT box with noise$^{\times}$	0.95	0.93	0.97	126.1	78.8	121.8	75.6	119.8	74.9	38.7	46.9	49.1	115.7	72.1	39.6	44.5	46.6
	AiOS box	0.95	0.93	0.97	124.3	78.7	120.0	75.7	118.1	74.8	37.4	45.5	47.0	114.0	71.9	38.3	43.2	44.7
SMPLer-X [2]	GT box$^{\times}$	0.95	0.93	0.97	101.5	61.4	97.6	58.8	96.4	58.3	31.7	40.8	41.9	92.7	55.9	32.6	38.3	39.4
	GT box with noise$^{\times}$	0.95	0.93	0.97	105.6	64.0	101.6	61.5	100.3	60.8	32.5	42.4	43.6	96.5	58.4	33.5	39.9	41.0
	AiOS box	0.95	0.93	0.97	103.3	63.5	99.6	61.1	98.1	60.3	31.3	40.4	41.8	94.6	58.0	32.3	38.0	39.4

Table 2. **AGORA validation set with noise bounding box.** OSX [8] and SMPLer-X [2] are finetuned on the AGORA training set. **GT box** means that this method uses the ground truth bounding box to crop the image. **GT box with noise** means translating the ground truth bounding box by **10%** of the image size in the horizontal direction, causing the human to deviate from the image center. This is a very small noise that ensures the person is not removed from the image plane, avoiding truncation. × means that this method filters the instances that AiOS fails to detect. The best results are colored with red, and the second-best results are colored with blue for OSX and SMPLer-X, respectively.

more instances with severe occlusion but resulting in many redundant detections. This is reflected in the metrics as high recall but lower accuracy.

From the data presented in Table 1, it's evident that regardless of using a low or high threshold for filtering detection results, both OSX [8] and SMPLer-X [2] show a marked decrease in performance in terms of NVME and NMJE when compared to results achieved using GT bounding boxes. This performance gap is largely due to the way NMVE and NMJE are normalized using the F1 score. The F1 score, being the harmonic mean of recall and precision, penalizes methods for both missed detections and false positives, which explains the observed discrepancy in performance.

Notably, when we filter the detection results with a score of 0.1, denoted by **Detected box *w* score 0.1**, the recall is 0.92, indicating that we detect the majority of instances, although there were many redundant detections. In this case, if we concentrate on reconstruction accuracy, represented by MVE and MPJPE, we observe that the results using detected bounding boxes OSX and SMPLer-X are significantly worse than using GT bounding box for cropping images.

On the lower part of Table 1, we present the AiOS results. To allow more hard cases to be detected, we use a threshold of 0.1 to filter the result. Similar to the main paper, providing AiOS's bounding boxes to OSX and SMPLer-X denoted by **AiOS box**, leads to a significant performance increase compared to using the detected boxes (**Detected box *w* score 0.3**), even though AiOS's box includes more challenging cases. Specifically, there is a significant improvement in NMVE with a 16% reduction from 148.1 to 124.3 for OSX and an 18% decrease from 126.4 to 103.3 for SMPLer-X. Additionally, there is an improvement in NMJE, with a 15% reduction from 142.1 to 120.0 for OSX and a 17% decrease from 121.3 to 99.6 for SMPLer-X. These results substantiate that one-stage methods demonstrate superior performance in real-world scenarios compared to existing two-stage methods.

For a fair comparison with OSX and SMPLer-X, which use the GT bounding boxes to crop images, we filter out the same instances that AiOS failed to detect for both OSX and SMPLer-X. Interestingly, we found that the results obtained using AiOS bounding boxes (denoted with **AiOS box** in Table 1) are comparable to those obtained using GT bounding boxes (denoted with **GT box**$^{\times}$). Notably, under this bounding box setting, AiOS still outperforms the current state-of-the-art OSX, even though it utilizes the GT bounding box and is on par with the foundational model SMPler-X L20. To further demonstrate the superiority of AiOS, we follow the procedure in RoboSMPLX [15] to translate the image by 0.1 of the image size horizontally, introducing a small noise to the GT bounding boxes (denoted by **GT box with noise**$^{\times}$ in Table 2). It's worth noting that this noise

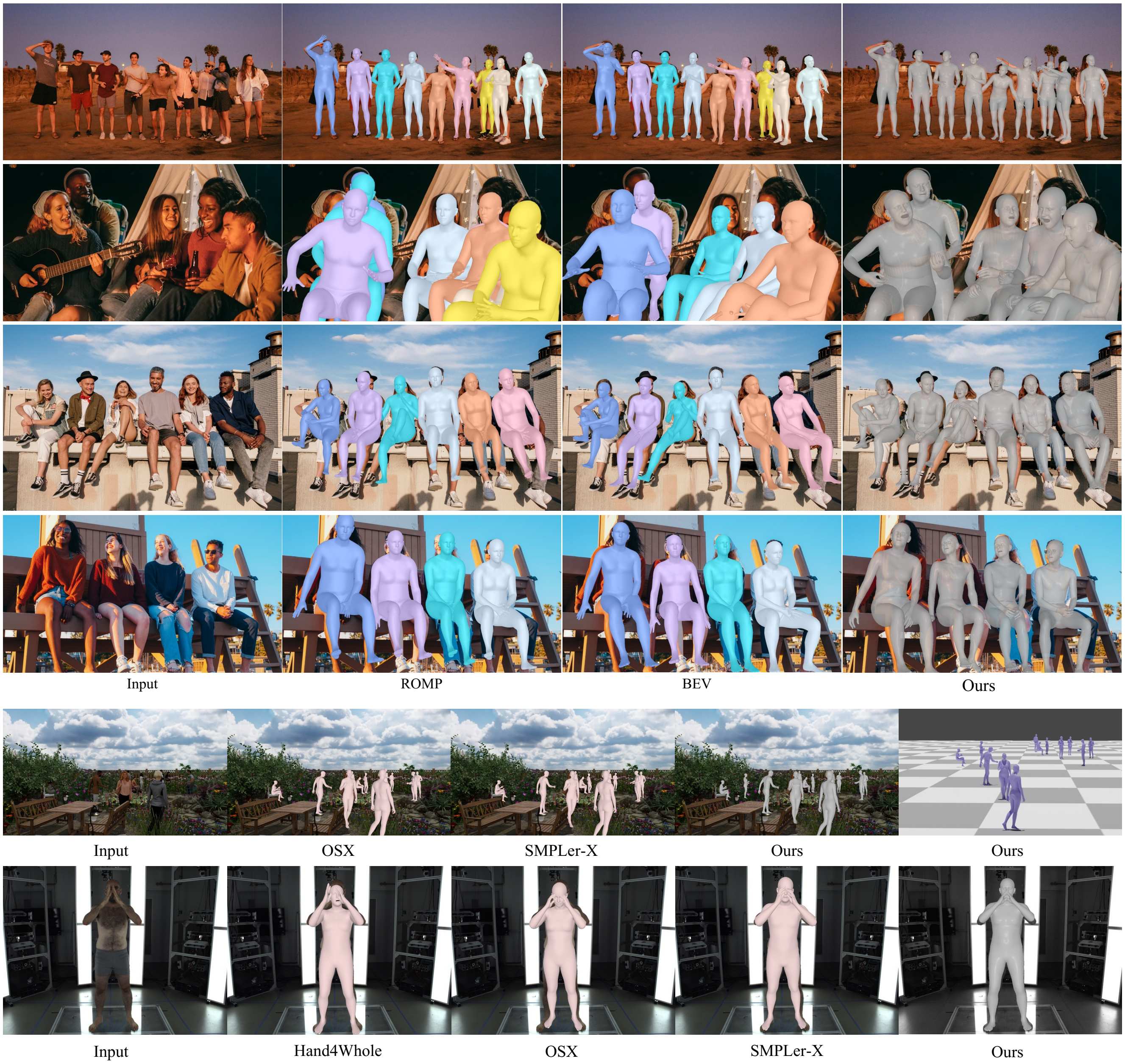

Figure 4. **Additional visual comparisons with existing methods. Upper part:** The first column is the input images, and they are downloaded from the internet. The second column is the visualization results of ROMP [19], the third column shows the visualization results of BEV [20], and the last column illustrates our visualization results; **Middle part:** Comparison of current SOTA methods [2, 8] with our AiOS model on AGORA [16]. **Lower part:**Comparison of current SOTA methods [2, 8, 13] with our AiOS model on EHF test [4].

Methods	F Score↑	Precision↑	Recall↑	NMVE↓ (*mm*)		NMJE↓ (*mm*)		MVE↓ (*mm*)					MPJPE↓ (*mm*)				
				All	Body	All	Body	All	Body	Face	LHand	RHand	All	Body	Face	LHand	RHhand
PIXIE [7]	0.94	0.99	**0.90**	158.7	107.2	153.7	103.5	149.2	100.8	51.4	44.8	48.9	144.5	97.3	55.4	41.3	44.8
BEDLAM-CLIFF	0.94	0.99	**0.90**	100.6	65.2	98.0	64.3	94.6	61.3	29.8	34.7	35.5	92.1	60.4	30.4	32.2	32.6
BEDLAM-CLIFF++†	0.94	0.99	**0.90**	93.2	61.2	90.9	60.4	87.6	57.5	27.3	30.3	32.6	85.4	56.8	28.0	28.0	29.9
AiOS†	**0.95**	**1**	**0.90**	**87.6**	**57.7**	**85.8**	**57.7**	**83.2**	**54.8**	**26.4**	**28.1**	**30.8**	**81.5**	**54.8**	**26.2**	**25.9**	**28.1**

Table 3. **BEDLAM test set.** The best results are in **bold**. † denotes methods that include the AGROA training set.

only shifts the person away from the image center and does not remove the person from the image plane, avoiding truncation and occlusion. Firstly, compare it with the results cropping with the GT bounding box denoted by **GT box**$^\times$,

	PA-PVE↓ (*mm*)			PVE↓ (*mm*)		
Method	All	Hands	Face	All	Hands	Face
H4W [13]	63.4	18.1	4.0	136.8	54.8	59.2
OSX [8]	56.9	**17.5**	3.9	102.6	56.5	44.6
OSX [8]†	33.0	18.8	3.3	58.4	39.4	30.4
SMPLer-X [2]	31.9	18.9	2.5	52.2	39.3	27.0
Native AiOS	31.8	19.7	2.3	51.6	40.6	26.5
AiOS	**30.2**	19.2	**2.1**	**47.1**	**38.3**	**26.1**

Table 4. **ARCTIC.** † denotes the method finetuned on the ARCTIC training set.

Method	PA-PVE↓ (*mm*)			PVE↓ (*mm*)		
	All	Hands	Face	All	Hands	Face
H4W [13]	58.8	9.7	3.7	121.9	50.0	42.5
OSX [8]	54.6	11.6	3.7	115.7	50.6	41.1
OSX [8]†	45.3	10.0	3.0	82.3	46.8	35.2
SMPLer-X [2]	38.9	9.9	3.0	66.6	42.7	31.8
SMPLer-X [2]†	**37.8**	9.9	**2.9**	63.6	46.3	32.3
Native AiOS	40.8	9.1	3.0	64.6	42.3	**26.3**
AiOS	38.0	**9.0**	**2.9**	**61.6**	**40.0**	26.7

Table 5. **EgoBody-EgoSet.** † denotes the methods that are finetuned on the EgoBody-EgoSet training set.

we can observe a drop in performance. Specifically, for NMVE, OSX increased from 123.6 to 126.1, and SMPLer-X increased from 101.5 to 105.6. Regarding NMJE, OSX increased from 119.3 to 121.8, and SMPLer-X increased from 97.6 to 101.6. This observation indicates that current two-stage methods are highly sensitive to the accuracy of the bounding box, as even a slight noise introduced, causing the person to be off-center, results in a performance drop, even with GT bounding boxes.

When comparing the results obtained by cropping images with bounding boxes provided by AiOS to those obtained by cropping with GT bounding boxes and GT bounding boxes with added noise, we observe that the results of cropping with AiOS-provided bounding boxes are slightly inferior to those obtained with GT bounding boxes in body-related metrics but better than GT bounding boxes with added noise. However, for the face and hands, the results of cropping with AiOS-provided bounding boxes can even be better than those obtained with GT bounding boxes. We attribute this improvement to AiOS's attention to not only the body but also the hands and face.

4. Extra SOTA comparison experiments

In this section, we show the extra single datasets on BEDLAM, ARCTIC, and EgoBody-EgoSet in Table 3, 4, 5. Our proposed AiOS achieved SOTA performance across all these datasets. Also, we provide extra visualization comparison in Fig. 4.

References

[1] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8726–8737, 2023. 1, 2

[2] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Yanjun Wang, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *arXiv preprint arXiv:2309.17448*, 2023. 1, 3, 4, 5, 6

[3] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 3

[4] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Monocular expressive body regression through body-driven attention. In *Eur. Conf. Comput. Vis.*, pages 20–40. Springer, 2020. 5

[5] MMHuman3D Contributors. Openmmlab 3d human parametric model toolbox and benchmark. `https://github.com/open-mmlab/mmhuman3d`, 2021. 1

[6] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12943–12954, 2023. 1, 2

[7] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael Black. Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision (3DV)*, pages 792–804, Dec. 2021. 5

[8] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 21159–21168, 2023. 1, 3, 4, 5, 6

[9] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis.*, pages 740–755. Springer, 2014. 1, 2, 3

[11] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In *Int. Conf. Learn. Represent.*, 2022. 3

[12] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):1–16, 2015. 1

[13] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2022. 5, 6

[14] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Neuralannot: Neural annotator for 3d human mesh training sets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2299–2307, 2022. 1

[15] Hui En Pang, Zhongang Cai, Lei Yang, Qingyi Tao, Zhonghua Wu, Tianwei Zhang, and Ziwei Liu. Towards robust and expressive whole-body human pose and shape estimation. In *Adv. Neural Inform. Process. Syst.*, 2023. 4

[16] Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. AGORA: Avatars in geography optimized for regression analysis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13468–13478, 2021. 1, 2, 3, 5

[17] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 1

[18] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019. 2

[19] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *Int. Conf. Comput. Vis.*, pages 11179–11188, 2021. 5

[20] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13243–13252, 2022. 5

[21] Jie Yang, Ailing Zeng, Shilong Liu, Feng Li, Ruimao Zhang, and Lei Zhang. Explicit box detection unifies end-to-end multi-person pose estimation. *arXiv preprint arXiv:2302.01593*, 2023. 1

[22] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *Eur. Conf. Comput. Vis.*, pages 180–200. Springer, 2022. 1, 2