REBAR: REFERENCE-BASED REASONING FOR ROBUST HUMAN POSE AND SHAPE ESTIMATION FROM MONOCULAR IMAGES
SUPPLEMENTARY MATERIALS

Anonymous authors
Paper under double-blind review

This supplement provides additional experimental results to enhance the main paper. In Section 1, we offer more implementation details. In Section 2, we present additional experimental results. In Section 3, we use more in-the-wild data to compare and verify the effectiveness of ReBaR on extremely challenging real-world problems.

1 IMPLEMENTATION DETAILS

In our experiments, our backbone network is initialized by the HRNet model weights pre-trained on the MPII (Andriluka et al., 2014) dataset for 2D key point detection tasks.

Part Segmentation Map
To obtain segmentation maps for the auxiliary supervision of the body attention map and part attention map in the AGE module, segmentation labels are required for each part of the original image. However, labeling the original image is a time-consuming and costly process. Luckily, we can utilize the SMPL model to obtain the vertex coordinates in the camera coordinate system that correspond to each image. Using weak perspective transformation techniques (Kissos et al., 2020), we can then generate the 2D vertex coordinates in the pixel coordinate system that are needed to generate the segmentation map. This method enables us to obtain the necessary segmentation maps without the need for costly and manual labeling of the original image.

Torso Plane and Relative Depth
As shown in figure 1, we select the shoulder, hip, and pelvis points on the human torso to establish the frontal, lateral, and transverse planes of the torso. The center points of the skin regions of each joint are defined as the actual joint positions, and joint axes are defined by bringing equations greater than 0 into the plane as positive directions, and the depth value is calculated based on the distance from the point to the plane. This is used as a basic fact label for relative depth, and explicit plane consistency constraints are constructed.

Loss Weight
In all experiments covered in this paper, we weight the loss of keypoints and SMPL parameters five times more than other losses.

Data Augmentation
In all the experiments in the paper, except that Table 2 does not use any augmentation on the Agora dataset, other experiments use the same data augmentation method as PARE.

Auxiliary Loss
In the process of training ReBaR, we respectively use global 2D/3D keypoints, body/part attention map and relative depth supervision to assist the network to learn body-aware part features. For the supervision of body/part attention, we only supervise on the COCO-EFT dataset, and reset the weight of L_{x_seg} in the loss function to 0 on the mix and 3DPW datasets.For the Agora dataset, we do not supervise the relative depth.

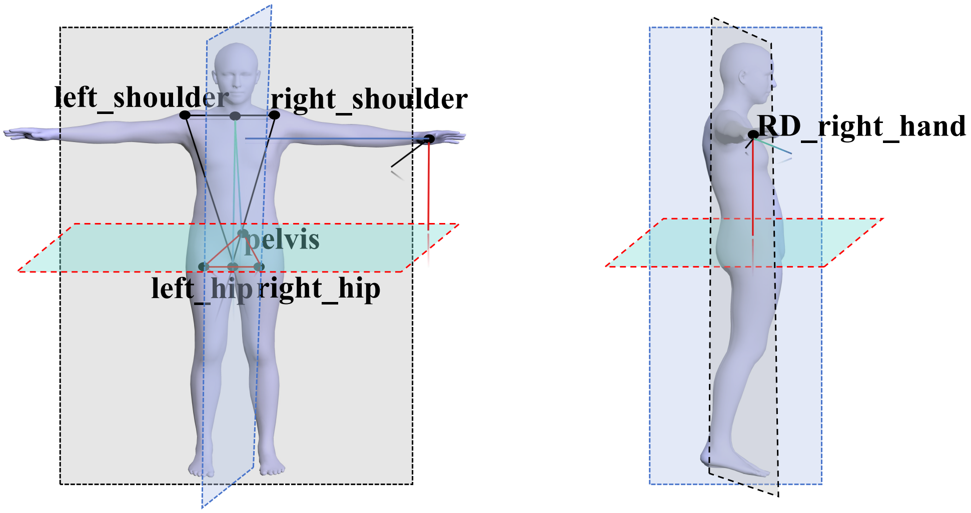

Figure 1: **Illustration of the calculation of the relative depth from the torso plane.** Shows the torso points for constructing the torso planes and the relative depth of the right hand.

2 EXPERIMENTS

In this section, we supplement more experimental results and ablation experiments to verify the effectiveness of ReBaR.

Method		Elbow			Wrist			Head	
	MJE$_x$↓	MJE$_y$↓	MJE$_z$↓	MJE$_x$↓	MJE$_y$↓	MJE$_z$↓	MJE$_x$↓	MJE$_y$↓	MJE$_z$↓
PARE	33.7	**27.5**	55.0	42.6	36.1	80.5	29.5	33.9	70.6
ReBaR	**27.9**	27.9	**48.5**	**36.1**	**35.5**	**75.6**	**20.5**	**30.0**	**48.1**
Method		Ankle			Knee			Neck	
	MJE$_x$↓	MJE$_y$↓	MJE$_z$↓	MJE$_x$↓	MJE$_y$↓	MJE$_z$↓	MJE$_x$↓	MJE$_y$↓	MJE$_z$↓
PARE	46.3	47.8	82.2	27.7	26.7	47.5	18.2	27.3	46.8
ReBaR	**35.0**	**36.6**	**71.0**	**21.8**	**17.9**	**40.4**	**15.3**	**24.1**	**36.0**

Table 1: **Part-by-part performance comparison on the 3DPW-Test.** All methods have been trained on dataset with 3DPW.

Per-part Evaluation
To validate the effectiveness of our method, we conducted more comprehensive evaluation experiments and reported the MPJPE metrics for each body part on each axis in Table 1. Compared to PARE, our method achieved a significant improvement of about 8mm on most parts, while the improvement of the shoulder in the z-axis is relatively small. Notably, our method achieved the most significant improvement on the Head, which dropped by 23.5mm in the z-axis. Some body parts such as the ankle and neck also showed substantial improvement, with a drop of 11.2mm and 10.8mm in the z-axis, respectively.

Analysis of BAR Module
BAR is the core part of ReBaR. Its role is to construct body perception component features and establish connections between components, so as to infer the posture of occluded parts through more reasonable visible information. This also makes ReBaR have a better ability to alleviate the depth blur problem than existing methods. As shown in Figure 2, we directly associate the component features of PARE and cannot correctly infer the correct posture of the occluded forearm (such as the wrong posture of the forearm bending forward), but after adding the body reference condition, the model infers Relatively correct posture, although its depth information is not very accurate, and then with the help of relative torso depth constraints, ReBaR accurately infers the posture of the arms behind it. In addition, we also found that in some extremely challenging actions, the body reference condition can greatly improve the stability of the root node. As shown in Figure 2, in the inverted

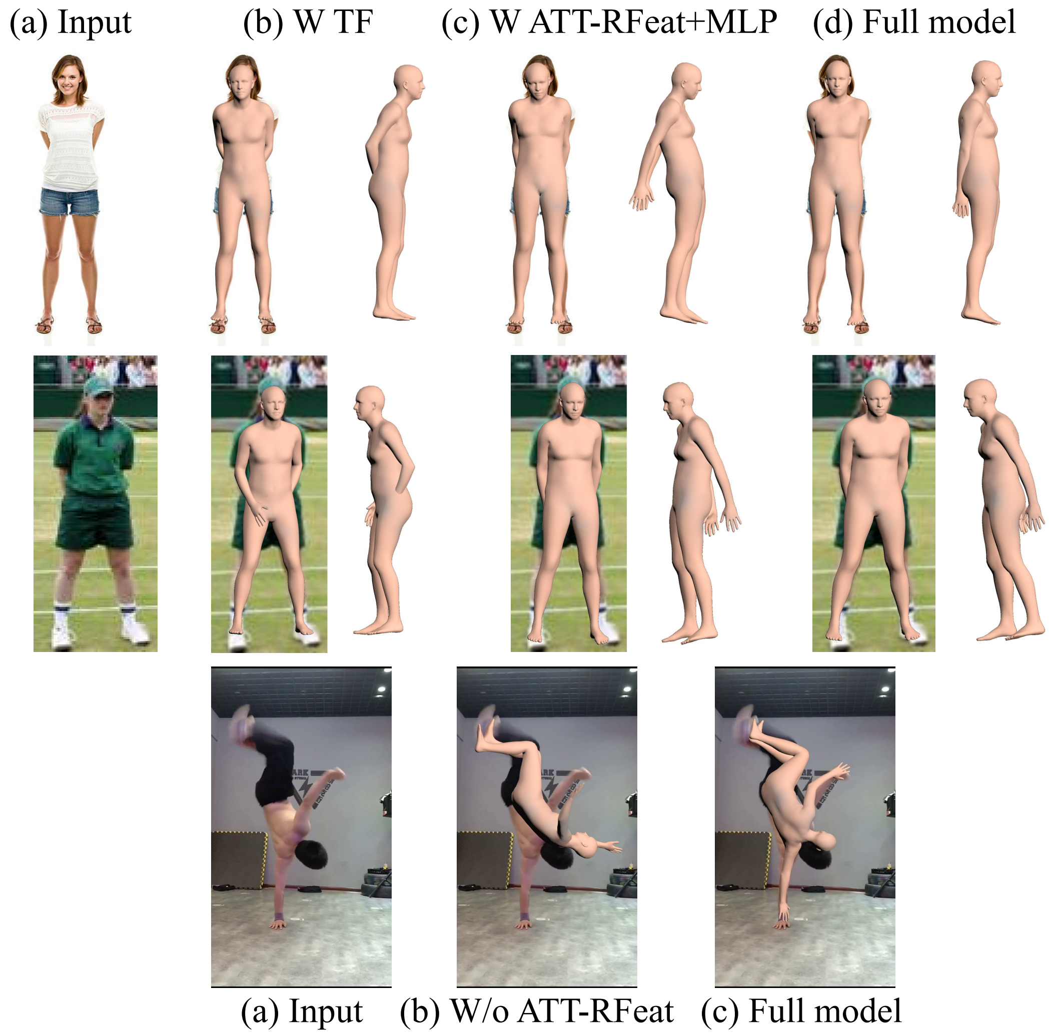

Figure 2: **The role of BAR module and ATT-RFeat.** For the first two rows, from left to right: input image, PARE+transformer, PARE+BAR, and the result of the full model. For the last line, from left to right: input image, w/o ATT-Rfeat, and the result of the full model.

Method	3DPW-All MJE↓	 PAMJE↓	 V2V↓
Baseline (Kocabas et al., 2021)	91.0	56.7	108.2
Baseline w TF	91.5	55.7	108.6
Baseline w HMR-RFeat+TF	90.1	55.0	106.1
ReBaR w/o L_{3D}	89.4	55.8	106.7
ReBaR w/o L_{2D}	90.6	54.8	107.2
ReBaR w/o L_{RD}	89.0	54.2	105.7
ReBaR	**88.6**	**53.7**	**105.3**

Table 2: **Ablation study of ReBaR on 3DPW.** All methods are trained on COCO-EFT-PART.

action, because PARE independently predicts the root joint, it causes an error in the global rotation direction. However, our method ReBaR accurately predicts a reasonable direction through the root feature of body perception, avoiding the problem of random rotation of the human body in video capture.

The Role of Attention-guided Reference Feature
Figure 3 illustrates the qualitative improvement of our method in challenging cases such as severe occlusion, challenging poses, and depth ambiguity. This demonstrates the importance of body-aware part features encoding and utilizing visual information around parts and Attention-guided reference feature to address depth ambiguity and self-occlusion issues.

Ablation Experiments With Attention-guided Reference Features
To evaluate the effectiveness of the body-aware regressor module, we conducted a set of comparative experiments. We used the global feature of the HMR model as a reference and directly associated it with the part feature to regress the SMPL model parameters using the same transformers. The results, as shown in Table 2, indicate that the global reference encoding only marginally improves the PAMJE metric compared to PARE. However, when we replaced the HMR feature with the Attention-guided reference feature of the AGE module, the model's performance significantly improved. This demonstrates the importance of using Attention-guided referenced feature for inferring part poses.

Ablation Experiments With Auxiliary Constraints
In this experiment, we use HR-W32 as the backbone and train all comparison methods on the small COCO-2014-EFT (22K) dataset (a subset of the COCO-EFT dataset). We first combine PARE with Transformers and evaluate the performance on the 3DPW dataset. Table 2 shows that integrating both techniques slightly reduces PAMJE but increases MJE and PVE. In contrast, our proposed ReBaR significantly improves PARE, indicating that learning Attention-guided reference feature substantially contributes to the performance gain. Furthermore, we validate the auxiliary constraints, i.e., 2D keypoints L_{2D}, 3D keypoints L_{3D}, and relative depth L_{RD}, which bring different levels of improvement to our proposed Attention-guided reference feature. Compared to the unconstrained HMR-feature and single-information supervised body-feature, establishing dependencies between 2D and 3D information in space can better construct stable reference conditions, thereby greatly improving joint prediction accuracy. The relative depth constraint provides greater weight on the depth axis and endows the model with the ability to perceive front-back relationships (positive outside the torso plane and negative otherwise), thereby alleviating the depth blur problem, which is an ability that pure 3D keypoint constraints do not possess. The results in Table 2 show that all losses contribute to improved performance.

Compared To Video-based Methods
We also compared our method to the state-of-the-art video-based methods. Table 3 shows the results, which demonstrate that our method outperforms these methods by a significant margin even without additional temporal information from the video.

3 MORE QUALITATIVE COMPARISONS

In this section, we provide more qualitative comparison results. Visualize PARE (Kocabas et al., 2021), CLIFF (Li et al., 2022) and ReBaR results on images and challenging video files respectively

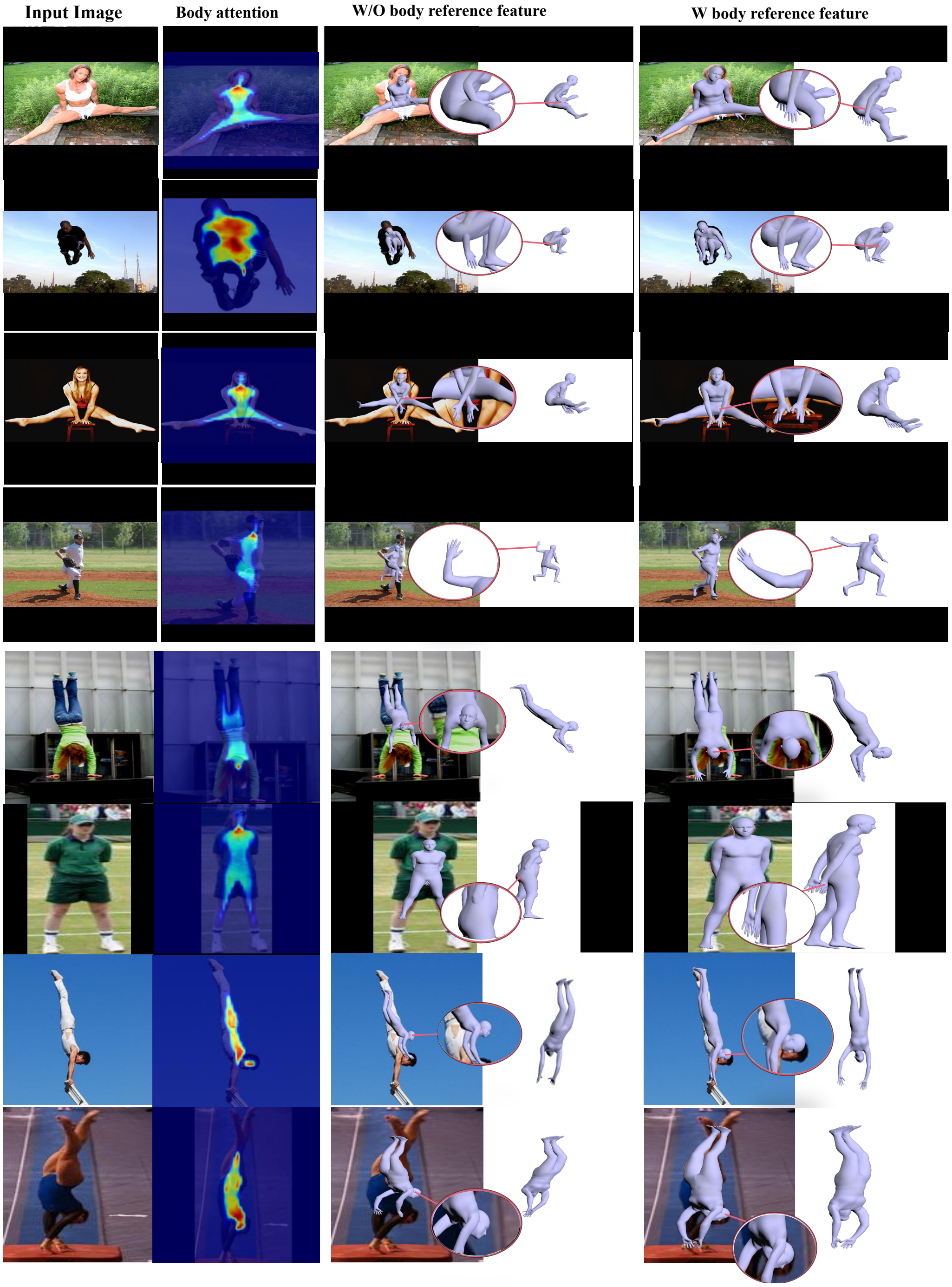

Figure 3: **The role of Attention-guided reference feature.** From left to right: input image, body attention map, the result of discarding Attention-guided reference feature, and the result of the full model.

Method	3DPW-Test	
	MJE↓	V2V↓
HMMR (Kanazawa et al., 2018)	116.5	72.6
Doersch et al. (Doersch & Zisserman, 2019)	-	74.7
Sun et al. (Sun et al., 2019)	-	69.5
TCMR (Choi et al., 2021)	105.3	100.7
VIBE (Kocabas et al., 2020)	82.7	51.9
MAED (Wan et al., 2021)	79.1	45.7
ReBaR	**69.1**	**41.8**

Table 3: **Evaluation on 3DPW-Test.** ReBaR achieves the best results without using the Euro filter."-" means no data provided.

for comparison. The image data includes two public datasets of 3DPW (Von Marcard et al., 2018) and LSPET (Johnson & Everingham, 2011), which contain problems such as occlusion, direction blur, and depth blur. Video files are downloaded directly from the internet and include challenging action sequences such as yoga, hip-hop and more.

Qualitative Comparison of Image As shown in Figure 4, our method outperforms PARE and CLIFF in almost all cases on the motion dataset LSPET. Especially under the problem of depth ambiguity and self-occlusion, thanks to the body-aware part features encoding, ReBaR can infer the spatial relationship between the occluded part and the body from the local visual cues around the part and the relevant global information of the whole body, Thereby improving the estimation accuracy. More interestingly, we found that PARE can barely determine the body orientation in some handstand situations, which we believe is due to PARE disconnecting the limbs and independently predicting the global rotation part. However, ReBaR avoids this problem nicely by using body reference conditions and limb dependencies.

Qualitative Comparison of Challenging Videos
As shown in Figure 5, we intercept some video frames for qualitative comparison. It is not difficult to find that in challenging action sequences such as hip-hop and yoga, the effects of PARE and CLIFF have dropped significantly, which shows that neither completely independent part prediction methods nor pure global prediction methods can handle actual complex movements in real-world applications. But ReBaR showed good results in handling these challenging actionse.

REFERENCES

Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pp. 3686–3693, 2014.

Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1964–1973, 2021.

Carl Doersch and Andrew Zisserman. Sim2real transfer learning for 3d human pose estimation: motion to the rescue. *Advances in Neural Information Processing Systems*, 32, 2019.

Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR 2011*, pp. 1465–1472. IEEE, 2011.

Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7122–7131, 2018.

Imry Kissos, Lior Fritz, Matan Goldman, Omer Meir, Eduard Oks, and Mark Kliger. Beyond weak perspective for monocular 3d human pose estimation. In *European Conference on Computer Vision*, pp. 541–554. Springer, 2020.

Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5253–5263, 2020.

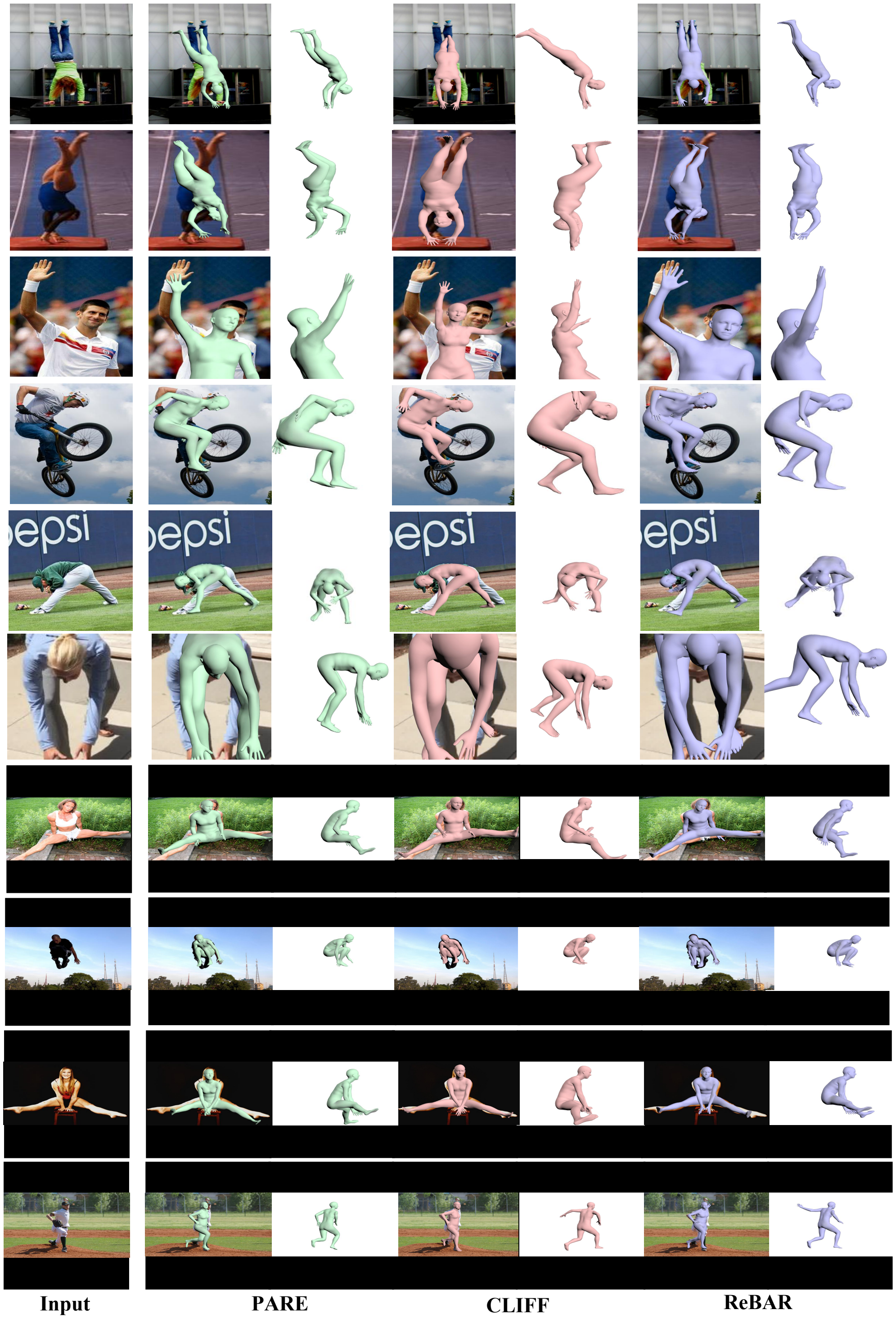

Figure 4: **Qualitative comparison of image.** From left to right: input image, PARE, CLIFF and ReBaR.

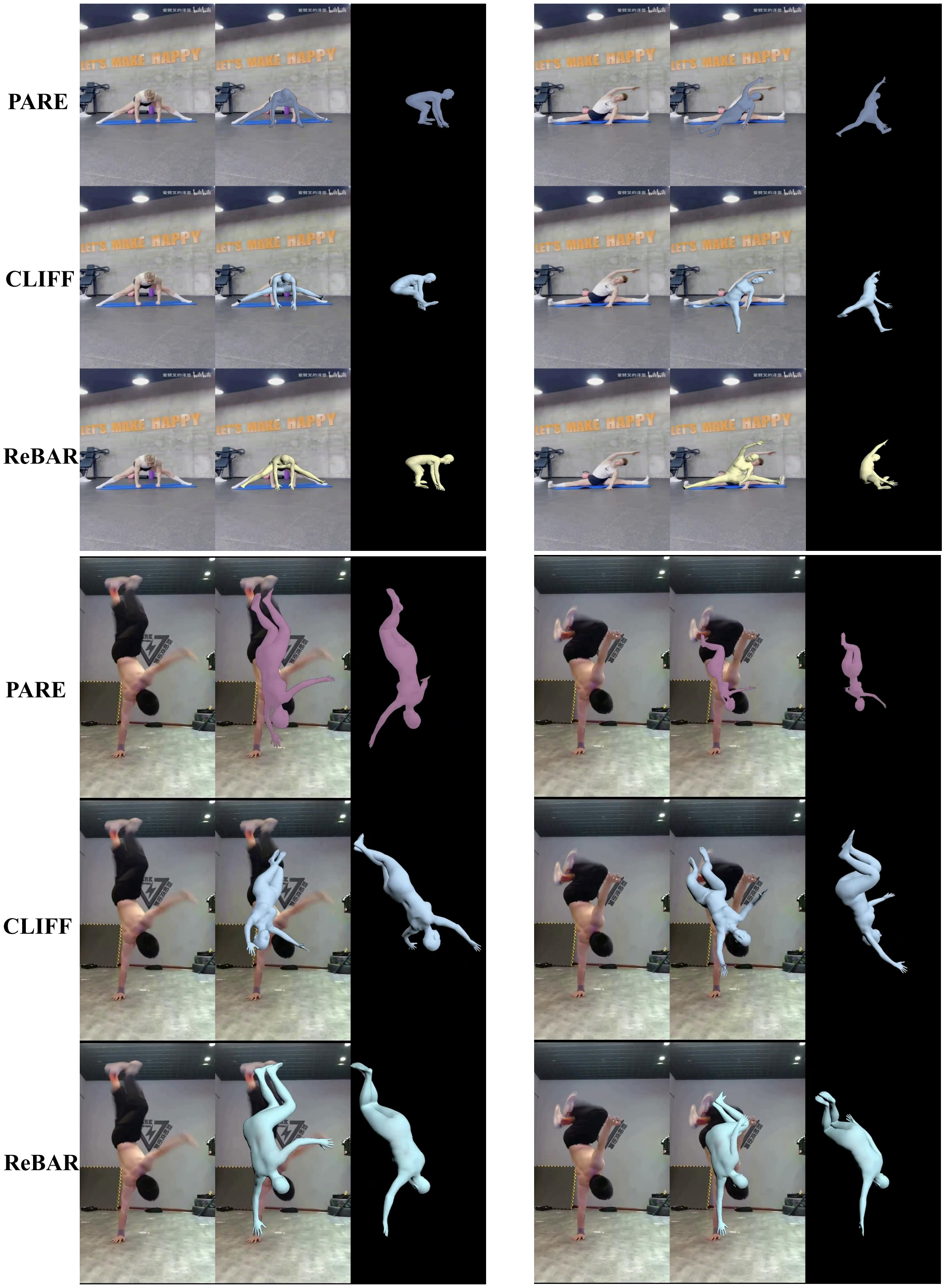

Figure 5: **Qualitative comparison of video.** From up to down: PARE, CLIFF and ReBaR.

Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11127–11137, 2021.

Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*, pp. 590–606. Springer, 2022.

Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5349–5358, 2019.

Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 601–617, 2018.

Ziniu Wan, Zhengjia Li, Maoqing Tian, Jianbo Liu, Shuai Yi, and Hongsheng Li. Encoder-decoder with multi-level attention for 3d human shape and pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13033–13042, 2021.