Towards Robust and Expressive Whole-body Human Pose and Shape Estimation

Hui En Pang[1], **Zhongang Cai**[1,2], **Lei Yang**[2], **Qingyi Tao**[2], **Zhonghua Wu**[2], **Tianwei Zhang**[1] ✉, **Ziwei Liu**[1]
[1]S-Lab, Nanyang Technological University [2]SenseTime Research
{huien001, tianwei.zhang, ziwei.liu}@ntu.edu.sg
{caizhongang, yanglei, taoqingyi, wuzhonghua}@sensetime.com

Abstract

Whole-body pose and shape estimation aims to jointly predict different behaviors (e.g., pose, hand gesture, facial expression) of the entire human body from a monocular image. Existing methods often exhibit degraded performance under the complexity of in-the-wild scenarios. We argue that the accuracy and reliability of these models are significantly affected by the quality of the predicted *bounding box*, e.g., the scale and alignment of body parts. The natural discrepancy between the ideal bounding box annotations and model detection results is particularly detrimental to the performance of whole-body pose and shape estimation. In this paper, we propose a novel framework RoboSMPLX to enhance the robustness of whole-body pose and shape estimation. RoboSMPLX incorporates three new modules to address the above challenges from three perspectives: **1) Localization Module** enhances the model's awareness of the subject's location and semantics within the image space. **2) Contrastive Feature Extraction Module** encourages the model to be invariant to robust augmentations by incorporating contrastive loss with dedicated positive samples. **3) Pixel Alignment Module** ensures the reprojected mesh from the predicted camera and body model parameters are accurate and pixel-aligned. We perform comprehensive experiments to demonstrate the effectiveness of RoboSMPLX on body, hands, face and whole-body benchmarks. Codebase is available at https://github.com/robosmplx/robosmplx.

1 Introduction

Human pose and shape estimation tries to build human body models from monocular RGB images or videos. It has gained widespread attention owing to its extensive applications in various fields, including robotics, computer graphics, and augmented/virtual reality. Early works use various statistical models (e.g., SMPL [31], MANO [43], FLAME [26]) to individually reconstruct different parts, including human body [17, 22, 4, 16, 24, 10, 21, 20], face [9, 8, 12], and hand [28, 3, 63]. Recently, there is a growing interest in whole-body estimation [11, 6, 61, 44, 58], which jointly estimates the pose, hand gestures and facial expressions of the entire human body from the input. Commonly these methods first employ separate sub-networks to extract the features of body, hands and face. These features are then used to predict whole-body 3D joint rotations and other parameters (e.g., body shape, facial expression), which are further combined to generate the whole-body 3D mesh. This is a crucial step towards modeling human behaviors in an efficient and practical manner.

However, achieving accurate and robust whole-body estimation is particularly challenging as it requires precise estimation of each body part and the correct connectivity between them. In particular, due to the smaller sizes of hand and face images, they are typically localized, cropped and resized to higher resolutions before being processed by the relevant sub-network. To tackle the absence

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

of ground-truth bounding boxes in the real-world scenarios, existing whole-body methods utilize various detection techniques to obtain the crops. The accuracy of the whole-body estimation is highly sensitive to the quality of input crops. Our experiment results in Section 3 show that even minor fluctuations in the scale and alignment of input crops can significantly affect the model performance, indicating a limited ability to localize and extract meaningful features about the subject in the image.

The lack of robustness in existing whole-body pose and shape estimation methods highlights three critical aspects that can be improved upon: 1) accurate localization of the subject and its parts, 2) accurate extraction of useful features, and 3) accurate pixel alignment of outputs. Inspired by these findings, we propose three novel modules, each specifically designed to address a particular goal:

- **Localization Module**. This module implements sparse and dense prediction branches to ensure the model is aware of the location and semantics of the subject's parts in the image. The learned location of the joint positions are helpful in recovering the relative rotations.
- **Contrastive Feature Extraction Module**. This module incorporates a pose- and shape-aware contrastive loss, along with positive samples, to promote better feature extraction under robust augmentations. By minimizing the contrastive loss, the model can produce consistent representations for the same subject, even when presented with different augmentations, making it robust to various transformations and capable of extracting meaningful invariant features.
- **Pixel Alignment Module**. This module applies differentiable rendering to ensure a more precise pixel alignment of the projected mesh, and learn more accurate pose, shape and camera parameters.

By integrating these three modules, we build a more robust and reliable whole-body pose and shape estimation framework, `RoboSMPLX`. Comprehensive evaluations demonstrate its effectiveness on body, face, hands and whole-body benchmarks.

2 Related Works

Whole-body Mesh Recovery. Despite significant progress in 3D body-specific [23, 22, 4, 16, 24, 10, 21, 20], hand-specific [28, 3], and face-specific [9] mesh recovery methods, there have been limited attempts to simultaneously recover all those parts. Early studies on whole-body pose and shape estimation primarily fit a 3D human model to 2D or 3D evidence [15, 53, 41, 54], which can be slow and susceptible to noise. Recent studies utilized neural networks to regress the SMPL-X parameters for a whole-body 3D human mesh. The model is composed of separate sub-networks to process body, hand and face, respectively. *One-stage* methods, e.g., OS-X [27], have the benefit of reduced computational costs and improved communication within part modules for more natural mesh articulation. However, the omission of hand and face experts makes it difficult for the model to leverage the widely available part-specific datasets, thus decreasing the hand and face performance. *Multi-stage* methods, e.g., ExPose [6], FrankMocap [44], PIXIE [11] and Hand4Whole [35], use different techniques to localize part crops.

Expose [41] and PIXIE [11] localize hand and part crops from the body mesh, making them dependent on the accuracy of body poses. Minor rotation errors accumulated along the kinematic chain may result in deviations in joint locations and thus inaccurate part crops. In contrast, Hand4Whole [35] predicts hand and face bounding boxes using a network leveraging image features and 3D joint heatmaps, but the resulting crops have low resolution. PyMAF-X [11] relies on an off-the-shelf whole-body pose estimation model to obtain crops, which, while more accurate, incurs extra computation. More detailed comparison with PyMAF-X are in Appendix C.

Robustness in vision tasks. Efforts to tackle robustness in vision tasks have utilized diverse strategies such as data augmentation, architectural innovations, and training methodologies [25, 56, 42, 30, 51, 60, 2]. AdvMix [51] employs adversarial augmentation and knowledge distillation, challenging models with corrupted images to foster learning from complex samples. Architectural modifications, such as novel heatmap regression [60], have been introduced to mitigate the impact of minor perturbations. HuMoR [42] utilizes a conditional variational autoencoder to capture the dynamics of human movement, thereby achieving generalization across diverse motions and body shapes. Additionally, PoseExaminer [30] employs a multi-agent reinforcement learning system to uncover failure modes inherent in human pose estimation models, highlighting model limitations in real-world scenarios. Complementing these efforts, Robo3D [25] provides a comprehensive benchmark for assessing the robustness of 3D detectors and segmentors in out-of-distribution scenarios.

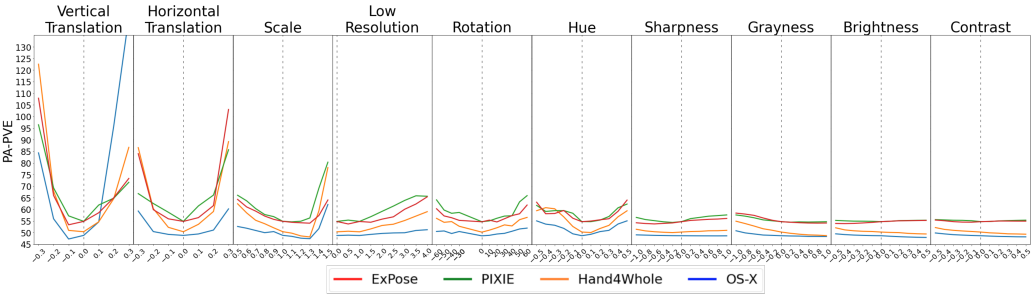

Figure 1: **Wholebody PA-PVE errors under different augmentations (sorted in descending order).** The dashed line indicates baseline performance without augmentation.

Furthermore, [56] utilize a confidence-guided framework to improve the accuracies of propagated labels. Contrastive learning, as demonstrated by CoKe [2], has also been employed to enhance robustness in keypoint detection, especially in occlusion-prone scenarios.

Contrastive Learning. Recently contrastive learning has demonstrated state-of-the-art performance among self-supervised learning (SSL) approaches. This strategy has been applied to 3D hand pose and shape estimation [46, 63]. Sanyal et al. [45] incorporate a novel shape consistency loss for 3D face shape and pose estimation that encourages the face shape parameters to be similar when the identity is the same and different for different people. Choi et al. [5] were the first to apply contrastive learning for 3D human pose and shape estimation. They found that SSL is not useful for this task, as the learned representations could be challenging to embed with high-level human-related information. Khosla et al. [19] proposed supervised contrastive learning for image classification tasks, which incorporates label information during training. Currently there is not attempt to apply this strategy to human pose and shape estimation, where the definition of positive samples is unclear, and data lie in a continuous space. We are the first to overcome these challenges and integrate supervised contrastive learning with whole-body pose and shape estimation.

Pixel Alignment in Pose and Shape Estimation. Many studies have been done to learn the subject's location in an image. Some works implicitly supervise the location. They primarily utilize projected meshes by supervising 2D joints regressed from the mesh [17, 23, 22, 4, 16, 24, 10, 21, 20]. Further supervision, such as dense body landmarks, silhouettes, and body part segmentation, is also employed to better align the predictions with the image [55, 37, 40, 49, 59, 57, 10]. Some other works explicitly learn the subject's location. Moon et al. [35] explicitly predict the keypoint locations in the image. Semantic body part segmentation is used as an explicit intermediate representation [41, 37]. PARE [41] employs a renderer to project the ground-truth mesh to the image space, and supervise the predicted part silhouette mask. However, dense part segmentation and differentiable rendering have not been employed in whole-body pose and shape estimation, which will be achieved in our framework.

3 Motivation

As discussed in Section 1, existing whole-body pose and shape estimation approaches suffer from the robustness issue, due to the models' sensitivity to the quality of input crops. To investigate the reasons and disclose the influence factors, we conduct a comprehensive evaluation of four state-of-the-art methods: ExPose [6], PIXIE [11], Hand4Whole [35] and OS-X [27]. We opt for a set of ten commonly encountered augmentations and vary their scales within a realistic range (see Appendix A for more details). The augmentations can be classed into three categories (1) *image-variant* augmentations: they affect the image without altering the objects' 3D poses or positions, such as color jittering; (2) *location-variant* augmentations: they modify the subject's location without changing its pose, involving operations like translation and scaling; (3) *pose-variant* augmentations: they simultaneously alter both the 3D pose and location, including rotation.

Impact of subject localization. We first reveal that existing models demonstrate high sensitivity to the subject's position, indicating potential difficulties in subject localization. Figure 1 reports the PA-PVE errors of the whole body under different augmentations. We observe that image-variant augmentations (contrast, sharpness, brightness, hue and grayscale) lead to an acceptable range of error rates (approximately in the 50s) and minimal fluctuation (around ± 2). In contrast, location-

Figure 3: **Sensitivity of existing body and hand models to different alignments (left) and scales (right).**

variant augmentations altering the subject's position within the frame, such as rotation, scaling, and horizontal or vertical translation, result in substantially higher error magnitudes. This demonstrates the heightened sensitivity of existing models to changes in the subject's position. In Appendix, we provide the results of other metrics and benchmarks in Figures 22 – 23, and visualizations of whole-body estimation under different settings in Figures 25 – 26.

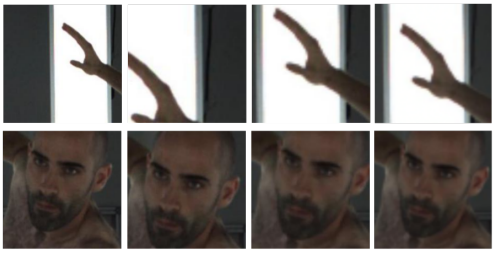

Figure 2: **Crops from (a) ExPose [6] (b) PIXIE [11], (c) Hand4Whole [35] (d)** `RoboSMPLX`**.**

Such position-altering augmentations are common in real-world scenarios, where the subject in the image is often localized using external detection models and control over the quality of crops is less feasible. In practice, to guarantee the visibility of the subject, crops are often made broader, This can lead to significant performance degradation, as errors increase with smaller augmentation scale factors (<1.0) (Figure 1). Besides, horizontal and vertical translations, which correspond to scenarios where the subject is not perfectly centralized or entirely visible within the frame, can further decrease the performance. Similarly, the alignment and scale of these crops also influence the pose and shape estimation systems targeting body, face and hands (Figure 3, more quantitative and qualitative evidence in Appendix M). Whole-body methods bear the additional responsibility of accurately localizing body parts such as hands and face. Inaccurate part crops (Figure 2) can adversely affect the performance of part subnetworks, and further the whole-body estimation.

Impact of feature extraction. The deterioration of performance in the face of such variations suggests that the model struggles to extract meaningful features. Under alterations in translation or scale, the subject remains within the image frame, though the proportion of background content may vary. It is difficult for existing methods to effectively disregard irrelevant background elements and extract relevant features related to the subject of interest. To enhance the model's robustness, it is critical to produce consistent features irrespective of various augmentations applied to the image.

Impact of output pixel alignment. Pixel alignment is a critical aspect of high model performance. In certain instances, despite having precise subject localization, the model fails to produce properly aligned results (Figure 25 in Appendix). This is often caused by the suboptimal camera parameter estimation. To address this issue, we need to accurately estimate the camera parameters, ensuring the projected mesh is precisely aligned with the ground-truth at the pixel level. Such precision would enhance the effectiveness of the model in producing accurate pose, shape and camera parameter predictions, improving the overall accuracy and reliability of the estimation process.

4 `RoboSMPLX` Framework

We design `RoboSMPLX` to enhance the robustness of whole-body pose and shape estimation. It provides three specialized modules to address each challenge in Section 3: 1) **Localization Module** (Section 4.2): explicitly learning the location information of the subject and incorporating it into model estimations for pose, shape and camera ; 2) **Contrastive Feature Extraction Module** (Section 4.3): reliably extracting pertinent features under various augmentations, thereby improving the model's generalization ability and robustness to a broader range of real-world scenarios; 3) **Pixel Alignment Module** (Section 4.4): ensuring that the outputs are pixel aligned.

We start with the description of `RoboSMPLX` architecture with Body, Hand and Face subnetworks (Section 4.1). Each subnetwork is integrated with the **Localization Module** and **Pixel Alignment**

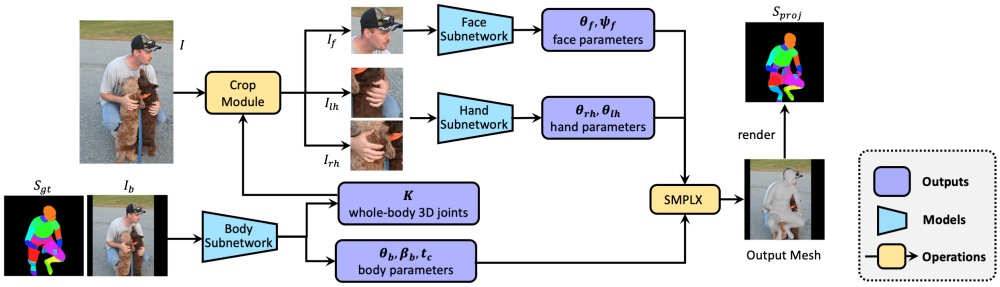

Figure 4: **Pipeline of our RoboSMPLX framework consisting of Body, Hand and Face subnetworks.**

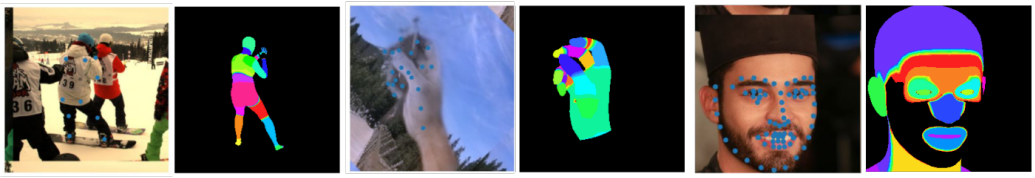

Figure 5: **Examples of keypoint and part segmentation supervision for Body, Hand and Face subnetworks.**

Module, and applies the **Contrastive Feature Extraction Module** for learning more robust features. Figure 6 shows the Hand subnetwork architecture. The other two subnetworks have the same designs.

4.1 Architecture and Training Details

Figure 4 shows the overall pipeline of RoboSMPLX for whole-body 3D human pose and mesh estimation. The Body subnetwork outputs 3D body joint rotations $\theta_b \in \mathbb{R}^{21\times3}$, global orientation $\theta_{bg} \in \mathbb{R}^3$, shape parameters $\beta_b \in \mathbb{R}^{10}$, camera parameters $\pi_b \in \mathbb{R}^3$, and whole-body joints $K \in \mathbb{R}^{137\times3}$. Joints corresponding to the hand and face are used to derive bounding boxes. Subsequently, hand and face images are cropped from a high-resolution image to preserve details. The Hand subnetwork predicts left and right hand 3D finger rotations $\theta_h \in \mathbb{R}^{15\times3}$. Simultaneously, the Face subnetwork generates 3D jaw rotation $\theta_f \in \mathbb{R}^3$ and expression $\psi_f \in \mathbb{R}^{10}$. When training Hand and Face subnetworks with part-specific datasets, additional parameters such as global orientation $\theta_{fg} \in \mathbb{R}^3$, shape $\beta_f \in \mathbb{R}^{50}$, and camera $\pi_f \in \mathbb{R}^3$ are estimated. These branches are discarded during whole-body estimation and training. Additional information concerning each subnetwork can be found in Appendix B. Further details regarding the training and inference durations are elaborated upon in Appendix K.

Subnetworks are trained separately, then integrated in a multi-stage manner. Initial whole-body training runs for 20 epochs. The hand and face modules are substituted with the trained Hand and Face subnetworks, followed by 20 epochs of fine-tuning to better unify the knowledge from the Hand and Face subnetworks into the whole-body understanding. Each subnetwork is trained by minimizing the following loss function L:

$$L = \lambda_{3D}L_{3D} + \lambda_{2D}L_{2D} + \lambda_{BM}L_{BM} + \lambda_{proj}L_{proj} + \lambda_{segm}L_{segm} + \lambda_{con}L_{con} \quad (1)$$

Here L_{BM} is the L1 distance between the predicted and ground-truth body model parameters. L_{3D} denotes the L1 distance between 3D keypoints and joints regressed from the body model. L_{2D} signifies the L1 distance of the ground-truth 2D keypoints to predicted and projected 2D joints. The latter are obtained by projecting the regressed 3D coordinates from the 3D mesh to the image space using the perspective projection [17]. The part segmentation loss L_{segm} is the cross-entropy loss between $P_{h,w}$ after softmax and $P_{h,w}$ averaged over H×W elements, following [20]. L_{proj} refers to the projected segmentation loss, which is the sigmoid loss between the projected mesh and the ground-truth segmentation map. L_{con} is the contrastive loss described in Section 4.3. For wholebody training, L_{box} is added to measure the L1 distance between the predicted and actual center and scale of the hands' and face's boxes.

4.2 Localization Module

This module focuses on subject localization by explicitly learning both sparse and dense predictions of the subject within the image. Figure 5 shows an example of the supervision used for each subnetwork.

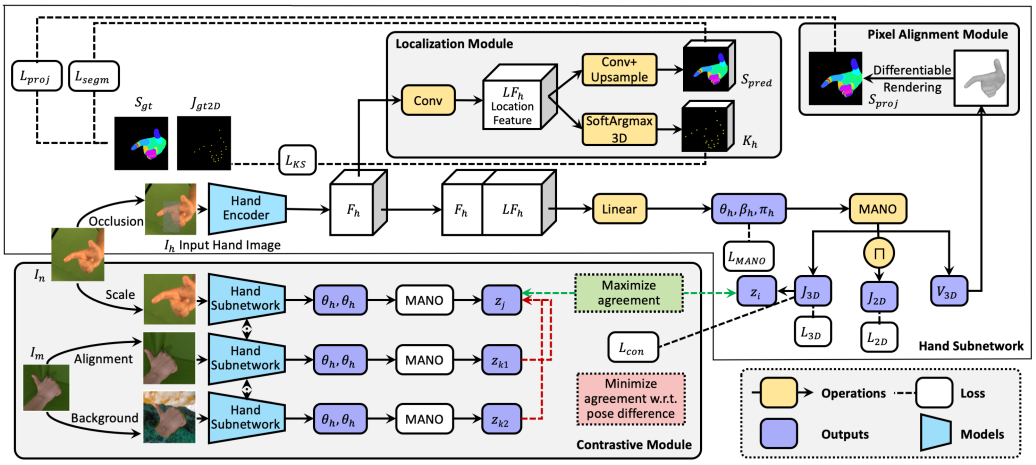

Figure 6: **Subnetwork Architecture with three modules.** We use the Hand subnetwork as an example. z represents normalized J_{3D} while p corresponds to the ground-truth of z. Green and red dashed lines refers to contrastive loss for positive and negative samples respectively.

Figure 7: **Augmentations for the Body subnetwork.** Black, blue and red labels represent image-variant, location-variant and pose-variant augmentations, respectively.

In contrast to prior methods that directly output pose rotations from backbone features, this module aims to make the model explicitly conscious of the subject's location and semantics while predicting pose, shape and camera parameters. It can reduce the model's sensitivity to the variations of the subject's position, caused by minor shifts in the scale and alignment of the bounding box.

As shown in Figure 6, given an image, a convolutional backbone is utilized to extract its feature map $F \in \mathbb{R}^{512\times32\times32}$. Following [35], a 1×1 convolutional layer is then used to predict 3D feature maps $LF \in \mathbb{R}^{32J\times32\times32}$ from F, where J represents the number of predicted joints with a feature map depth of 32. LF contains valuable information about the mesh's position in the image and semantics of various parts. It is concatenated with the backbone feature map F to predict pose $\theta \in \mathbb{R}^{P}$, shape $\beta \in \mathbb{R}^{10}$ and camera translation $\pi \in \mathbb{R}^{3}$, where P is the number of body parts. Meanwhile, LF is also used to obtain extra information with two branches: (1) 3D joint coordinates $K \in \mathbb{R}^{J\times3}$ are obtained from LF using the soft-argmax operation [47] in a differentiable manner. (2) 2D part segmentation maps $S \in \mathbb{R}^{P+1\times64\times64}$ are extracted from LF with several convolution layers, which model P part segmentation and 1 background mask. Here, 64 represents the height and width of the feature volume, and each pixel (h, w) stores the likelihood of belonging to a body part P.

Note that learning part segmentation maps and 3D joint coordinates is complementary, as 3D joint coordinates encode depth information that may inform part ordering in segmentation maps. Additionally, joints often reside at the boundaries of part segmentation maps, serving as separators for distinct parts. The Body subnetwork utilizes 24 parts P and 137 joints J, the Hand subnetwork employs 16 parts P and 21 joints J, while the Face subnetwork employs 15 parts P and 73 joints J.

4.3 Contrastive Feature Extraction Module

This module incorporates a pose- and shape-aware contrastive loss, along with positive samples. By minimizing this loss, the model can produce consistent representations for the same subject, even when presented with different augmentations, thus fostering the extraction of meaningful features.

Conventional contrastive learning methods based on SSL (e.g., SimCLR) face challenges in unifying similar pose embeddings and distancing dissimilar ones in human pose and shape estimation tasks. Without labels for guidance, images with similar poses could be misidentified as negative samples and contrasted away, complicating the self-organization of the embeddings in pose space. Figures 9 to 12 in Appendix show their ineffectiveness for the 3D human pose and shape task [5] by visualizing the retrieved samples from the embeddings. The supervised contrastive learning approach by Khosla

et al. [19], though effective for image classification, might not extend well to human pose and shape estimation, which is a high-dimensional regression problem and poses exist in a continuous space rather than well-defined classes.

Our module overcomes the aforementioned issues with two innovations. First, we experiment with three human pose representations $\boldsymbol{z}$ and the corresponding distance functions: (1) A concatenated form of the global orientation and rotational pose; (2) global orientation and rotational pose as separate entities (3) 3D root-aligned joints regressed from the body model, derived from pose and shape inputs. For (1) and (2), we explore relative rotations in two forms: 6D vector and rotation matrix representation. For (3), L1, Smooth L1, and Mean Squared Error (MSE) was used (Table 9).

Second, we investigate ten data augmentations, and classify them into three categories (see Figure 7 for the Body subnetwork, and Figure 29 in Appendix for Hand subnetwork): (1) *image-variant* augmentations such as color jittering, blur, occlusion and background swapping; (2) *location-variant* augmentations involving translation and scaling; (3) *pose-variant* augmentations including rotation and horizontal flipping. Our ablation study in Table 10 shows that augmentations with varied global orientation are detrimental to the model performance. Consequently, we exclude such modifications when constructing positive pairs. Instead, each positive sample is constructed utilizing a random combination of location-variant and color-variant augmentations.

Formally, for a batch of N images, we construct another N images by applying augmentation to each sample. For each anchor i, let j be the corresponding augmented sample. Then i is contrasted against $2N-1$ terms (1 positive and $2N-2$ negatives). The loss takes the following form:

$$\mathcal{L}_{con} = \sum_{i=1}^{N} \left(\tau_{pos} \left(\left| \mathrm{d}\left(\boldsymbol{p}_i, \boldsymbol{p}_j\right) - \mathrm{d}\left(\boldsymbol{z}_i, \boldsymbol{z}_j\right) \right| \right) + \tau_{neg} \sum_{k=1}^{2N} \mathbb{1}_{[k \neq i,j]} \left(\left| \mathrm{d}\left(\boldsymbol{p}_i, \boldsymbol{p}_k\right) - \mathrm{d}\left(\boldsymbol{z}_i, \boldsymbol{z}_k\right) \right| \right) \right) \tag{2}$$

where $\boldsymbol{z}_i$, $\boldsymbol{z}_j$ and $\boldsymbol{z}_k$ denote the predicted pose representations, and $\boldsymbol{p}_i$, $\boldsymbol{p}_j$ and $\boldsymbol{p}_k$ denote the ground-truth pose representations for the anchor, positive and negative samples in the batch. The objective of this loss function is to minimize the distance between the positive pairs and maximize the distance between the negative pairs, in alignment with the pose similarity. Note that unlike traditional approaches where the distance is the same for all negative samples, the pairwise distance $\mathrm{d}(\boldsymbol{p}_i, \boldsymbol{p}_k)$ varies depending on the pose similarity.

4.4 Pixel Alignment Module

This module employs differentiable rendering to ensure that the projected mesh aligns precisely at the pixel level. The alignment is supervised by the projected mask loss. Attaining a proper alignment between the ground-truth part segmentation and rendered mesh requires the accurate prediction of pose, shape, and camera parameters, which subsequently leads to a more precise estimation process.

5 Experiments

Datasets. For whole-body training, we employ Human3.6M (H36M) [13], COCO-Wholebody [14] (the whole-body version of MSCOCO [29]) and MPII [1]. The 3D pseudo-ground truths for training are acquired using NeuralAnnot [36]. For hand-specific training, we use FreiHAND [62], Interhand [34] and COCO-Wholebody Hands [14]. For face-specific training, we use FFHQ [18], BUPT [52] and AffectNet [32]. For evaluations specific to 3D body, 3D hand, and 3D face, we utilize 3DPW [50], FreiHAND [62], and Stirling [11], respectively. For the 3D whole-body evaluation, we use EHF [41] and AGORA [39]. Additionally, we present qualitative results on the MSCOCO validation set.

Metrics. Mean Per Joint Position Error (MPJPE) and Mean Per-Vertex Position Error (MPVPE) are employed to evaluate the positions of 3D joint and mesh vertices, respectively. Each metric calculates the average 3D joint distance (in *mm*) and 3D mesh vertex distance (in *mm*) between the predicted and ground-truth values after aligning the root joint translation. The pelvis serves as the root joint for whole-body and body, whereas the wrists and neck are utilized as root joints for hands and face. Procrustes Aligned (PA) variants of these metrics, PA-MPJPE and PA-MPVPE, further align with rotation and scale. We report the average errors for the left and right hands as the 3D hand error.

Table 1: **Evaluation of the Hand subnetwork**.

Method	PA-PVE ↓	PA-MPJPE ↓	F-Scores ↑
*** Hand-only**			
FreiHAND [62]	10.7	-	0.529/0.935
Pose2Mesh [4]	7.8	7.7	0.674/0.969
I2L-MeshNet [33]	7.6	7.4	0.681/0.973
METRO (HR64) [28]	**6.7**	**6.8**	0.717/0.981
*** Whole-body**			
ExPose [41]	11.8	12.2	0.484/0.918
Zhou et al. [61]	-	15.7	-/-
FrankMocap [44]	11.6	9.2	0.553/0.951
PIXIE [11]	12.1	12	0.468/0.919
Hand4Whole † [35]	7.7	7.7	0.664/0.971
HMR (Baseline) [17]	8.6	8.9	0.605/0.963
PyMAF [58]	8.1	8.4	0.638/0.969
PyMAF † [58]	7.5	7.7	0.671/0.974
RoboSMPLX	**7.3**	**7.5**	0.683/0.976
RoboSMPLX †	**7.1**	**7.4**	0.688/0.978
RoboSMPLX **(HR64)**	**6.7**	**6.9**	0.715/0.981

Table 2: **Evaluation of the Body subnetwork**.

Method	PA-MPJPE ↓	MPJPE ↓	PVE ↓
HMR (Res50) [17]	76.7	130	-
GraphCMR (Res50) [24]	70.2	-	-
SPIN (Res50) [22]	59.2	96.9	116.4
HMR-EFT (Res50) [16]	54.3	-	-
ROMP (Res50)	53.5	89.3	105.6
PARE (Res50) [20]	52.3	82.9	99.7
PARE (HR32) [20]	50.9	82	97.9
PyMAF (Res50)[58]	49.0	79.7	94.4
PyMAF (HR48) [58]	**47.1**	**78.0**	91.3
Baseline (Res50)	52.4	85.2	103.6
RoboSMPLX (Res50)	49.8	80.8	96.7
Baseline (HR48)	50.3	84.5	101.5
RoboSMPLX (HR48)	48.5	80.1	95.2

Table 3: **Evaluation of the Face subnetwork**.

Method	LQ Mean(mm) ↓	HQ Mean(mm) ↓
ExPose [6]	2.27	2.42
ExPose †	2.46	2.38
HMR	2.18	2.11
HMR †	2.31	2.27
HMR *	2.02	2.04
PyMAF *	1.97	1.92
RoboSMPLX	**2.12**	**2.08**
RoboSMPLX †	**2.12**	**2.10**

Table 4: **PA-PVE/PVE errors of the Hand subnetwork under different positional augmentations.**

	Normal	Transx +0.2x	Transx -0.2x	Transy +0.2y	Transy -0.2y	Scale 1.3x	Scale 0.7x
Hand4Whole [35]	7.47/ 15.70	8.51/ 21.58	8.38/ 20.36	8.74/ 22.51	8.48/ 19.85	7.73/ 16.44	7.78/ 17.00
RoboSMPLX	**7.24/ 15.23**	**7.27/ 15.62**	**7.36/ 15.59**	**7.28/ 15.50**	**7.34/ 15.50**	**7.49/ 15.90**	**7.45/ 16.51**

5.1 Benchmarking Results

Hand Subnetwork. Table 1 compares the performance of the Hand subnetwork with different hand-only and whole-body methods. Our method outperforms that of our whole-body counterparts when trained with only the FreiHAND dataset (i.e. PIXIE, Hand4Whole, PyMAF) or under mixed datasets (i.e. Hand4Whole †, PyMAF †)[1] using an identical backbone. Prior research [33, 48] demonstrated that whole-body methods generally employ a parametric representation of the hand mesh, and are numerically inferior to the non-parametric representation used in recent hand-only methods [33, 28]. Despite such reported gap, RoboSMPLX manages to outperform mesh-based techniques, and achieve comparable results as the state-of-the-art METRO when using the same backbone (HRNet-64). Table 4 compares the estimation errors of the Hand subnetwork in Hand4Whole (current whole-body method with SOTA on hands) and RoboSMPLX under different positional augmentations on the FreiHAND test set. It is clear that RoboSMPLX exhibits much better robustness than Hand4Whole. More visualizations are provided in Figure 20 in Appendix.

Body Subnetwork. Table 2 compares the performance of the Body subnetwork across different methods on the 3DPW test set. We observe the competitiveness of RoboSMPLX in relation to other SMPL-based approaches. Besides, since the performance of various methods may significantly differ based on their backbone initialization, datasets and training strategies [38], we establish a baseline to evaluate the effectiveness of our added modules in Table 12 in Appendix. RoboSMPLX achieves a substantial improvement compared to the baseline.

Face Subnetwork. Table 3 compares the performance of the Face subnetwork for different methods on the Stirling3D test set. When training with the same dataset, RoboSMPLX outperforms ExPose. The performance of ExPose declines when training on multiple datasets, while RoboSMPLX can still keep low and consistent errors. Figure 8 in Appendix shows some qualitative results for the in-the-wild scenarios, which demonstrates the high generalization of RoboSMPLX. Table 5 compares the robustness of ExPose and RoboSMPLX under different positional augmentations. We also observe that RoboSMPLX has lower errors with different translation and scaling operations. More visualizations are provided in Figure 21 in Appendix.

Whole-body Network. We further provide results of the whole-body network on two benchmarks: EHF val set and AGORA test set in Table 6. On EHF, RoboSMPLX outperforms other full-body approaches, particularly in hand and face performance evaluations, and under different positional augmentations (Table 7). It gives subpar performance on AGORA as the predominant source of

[1] † denotes training with extra datasets in the following evaluation and tables.

Table 5: **3DRMSE errors of the Face subnetwork under different positional augmentations.**

	Normal	Transx +0.2x	Transx -0.2x	Transy +0.2y	Transy -0.2y	Scale 1.3x	Scale 0.7x
ExPose [6]	2.27	2.38	2.29	2.46	2.30	2.46	2.27
RoboSMPLX	**2.12**	**2.20**	**2.17**	**2.13**	**2.18**	**2.24**	**2.10**

Table 6: **Evaluation of wholebody network on EHF and AGORA test set.**

Method	EHF						AGORA					
	PVE ↓			PA-PVE ↓			PVE ↓				N-PVE ↓	
	WB	H	F	WB	H	F	WB	B	F	LH/RH	WB	B
ExPose [6]	77.1	51.6	35	54.5	12.8	5.8	217.3	151.5	51.1	74.9/71.3	265	184.8
PIXIE [11]	89.2	42.8	32.7	55	11.1	**4.6**	191.8	142.2	50.2	49.5/49.0	233.9	173.4
Hand4Whole [35]	76.8	39.8	26.1	50.3	10.8	5.8	135.5	90.2	41.6	46.3/48.1	144.1	96.0
OSX (ViT-L)	70.8	53.7	26.4	**48.7**	15.9	6.0	**122.8**	**80.2**	**36.2**	45.4/46.1	**130.6**	**85.3**
PyMAF-X (HR48)	**64.9**	**29.7**	19.7	50.2	10.2	5.5	125.7	84	35	**44.6/45.6**	141.2	94.4
Ours	73.7	34.9	**17.8**	49.7	**10.0**	**4.6**	132.3	85	39.4	45.3/46.1	138.2	91.5

error is the misidentification of individuals under intense person-person occlusion. We give detailed investigation in Appendix D.

5.2 Ablation Studies

Contrastive loss. We validate prior contrastive SSL methods [63, 46, 5] are not particularly adept at learning useful embeddings for human pose and shape estimation. Figures 9 – 12 in Appendix visualize the retrieved images based on the top-5 embedding similarity. They show that without labels, the model primarily extracts features based on background information instead of pose information. Table 8 shows the estimation errors of top-1 retrieved pose (COCO-train) and query pose (COCO-test) with different methods and contrastive loss functions. We observe that SimCLR has higher mean errors than the supervised training method HMR. These results are aligned with [5] that the representations learned through SSL are not transferable for human pose and shape estimation tasks. RoboSMPLX incorporates contrastive loss and positive samples ("HMR + L_{con}, +ve"), which can produce similar representations under varied augmentations, enhancing its robustness.

Table 9 shows the estimation errors when applying contrastive loss with different representations in Section 4.3: "pose" (a concatenated form of global orientation and rotational pose), "go+pose" (global orientation and rotational pose as separate entities), "keypoint" (3D joints regressed from the body model). We observe that regressed 3D joints are the most effective representation, as they encode both shape and pose information in a normalized space. In contrast, the representation of pose as relative rotation has a detrimental impact on the model performance. Incorporating positive samples ("pose, +ve" and "keypoint, +ve") bolsters contrastive learning, encouraging the model to generate similar representations under varied augmentations. Table 10 compares the model performance with different augmentations. Prior methods [63, 46] employed pose-variant augmentations (e.g., rotation and flipping), which can adversely affect the learning by altering the global orientation, and lead to increased errors ("pose") compared to "baseline". Conversely, color-variant, location-variant and their combination provide an improvement over the baseline, showing these augmentations are helpful.

Location features. Table 11 shows the ablation of different modules on the Hand subnetwork (ablation for the Body subnetwork is in Table 12 in Appendix). The baseline model is trained that randomly augments images with a scale factor of 0.2 and bounding box jitter of 0.2. We observe that training using *strongaug* with a larger scale and jitter factor harms the baseline performance. This is likely due to a domain shift. Hand4Whole [35] employs sampled features from positional pose-guided pooling (PPP) to predict pose parameters while shape and camera parameters only utilize backbone features. Our method focuses on explicitly learning the location and part silhouettes, utilizing sparse and dense supervision methods. This proves advantageous as the location information ("LF") improve the performance of pose and shape estimations, with the reduced joint and vertex errors of the regressed mesh. Moreover,

Table 11: **Ablation of different modules on Hand subnetwork. Results are trained and evaluated on FreiHAND.**

	Supervision	PA-↓	MPJPE↓	PA-↓	PVE↓
Base (R50)		8.06	16.78	7.85	16.71
Base (R50) + Strongaug		8.47	17.01	8.11	16.17
Base (DR54)		7.8	15.57	7.67	15.72
Base (DR54)	L_{KS}	7.68	15.8	7.62	16.29
PPP [35]	L_{KS}	7.65	15.93	7.56	16.37
LF	L_{KS}	7.52	15.84	7.56	16.15
joints	L_{KS}	7.86	15.92	7.75	16.24
LF (all)	L_{KS}	7.49	15.51	7.46	15.59
LF (all) + L_{con}	L_{KS}	7.48	15.01	7.32	15.29
LF (all) + L_{con}, +ve	L_{KS}	7.42	14.88	7.16	14.57
LF (all)	L_{KS}, L_{segm}	7.44	14.92	7.58	15.30
LF (all)	L_{KS}, L_{segm}, L_{proj}	7.36	**14.38**	7.53	15.05
LF (all)+ L_{con}, +ve	L_{KS}, L_{segm}, L_{proj}	**7.33**	14.59	**7.02**	**14.11**

Table 7: **Wholebody, Hand and Face PA-PVE errors under different positional augmentations.**

	Method	Normal	Transx +0.2x	Transx -0.2x	Transy +0.2y	Transy -0.2y	Scale 1.3x	Scale 0.7x
Hands	ExPose [6]	14.39	17.36	17.86	14.93	17.21	14.15	14.56
	PIXIE [11]	14.68	15.05	16.11	15.32	15.85	14.52	14.79
	Hand4Whole [35]	10.83	11.15	11.34	10.50	13.70	10.77	11.25
	OSX [27]	15.97	16.42	16.55	16.94	17.86	15.91	17.24
	RoboSMPLX	**10.00**	**10.37**	**10.21**	**10.16**	**12.49**	**9.98**	**10.19**
Face	ExPose [6]	6.34	10.28	6.71	8.17	6.43	6.24	6.24
	PIXIE [11]	5.63	6.67	6.94	6.53	6.94	5.84	5.84
	Hand4Whole [35]	5.81	5.88	5.91	5.74	5.93	5.76	5.76
	OSX [27]	6.09	6.03	6.09	5.83	5.96	5.92	5.92
	RoboSMPLX	**4.65**	**5.10**	**5.38**	**4.75**	**5.30**	**4.77**	**5.22**
Wholebody	ExPose [6]	54.82	61.64	65.98	65.03	65.98	54.03	59.23
	PIXIE [11]	54.85	66.16	69.26	64.83	69.26	56.28	60.31
	Hand4Whole [35]	50.37	59.10	67.85	64.64	67.85	48.10	55.28
	OSX [27]	**48.79**	**51.09**	55.96	95.97	**55.96**	**47.35**	**50.89**
	RoboSMPLX	49.79	52.46	**53.62**	**61.65**	63.99	47.90	51.39

Table 8: **Ablation of contrastive learning methods and loss.**

Scale factor	Mean ↓	Std ↓
SimCLR	0.227	0.0915
SimCLR (+ pose-variant aug.)	0.230	0.0911
SimCLR (+ background aug.)	0.222	0.0959
SimCLR (+ L_{con})	0.164	0.0772
HMR	0.140	0.0823
HMR (+ L_{con})	0.124	**0.0624**
HMR (+ L_{con}, +ve samples)	**0.119**	0.0679

Table 9: **Ablation of different representation for contrastive loss.**

Representation	PA-↓	MPJPE↓	PA ↓	PVE↓
baseline	7.49	15.51	7.46	15.59
pose	8.11	15.81	7.67	16.08
go + pose	7.71	14.98	7.54	14.91
keypoint	7.48	15.01	7.32	15.29
pose, +ve	7.45	14.94	7.20	**14.77**
keypoint, +ve	**7.31**	**14.62**	**7.18**	15.01

Table 10: **Ablation of augmentation +ve samples, using pose rotation as representation.**

Augmentation	PA-↓	MPJPE↓	PA- ↓	PVE↓
baseline (no +ve)	8.11	15.81	7.67	16.08
color	**7.42**	15.01	**7.18**	14.94
pose	8.59	16.96	8.15	17.21
location	7.80	15.98	7.46	15.56
color + location	7.45	**14.94**	7.20	**14.77**

we find that using location features "LF (all)" for predicting shape and camera parameters is also beneficial.

Pixel alignment. Tables 11 also shows that incorporating differential rendering and using projected segmentation loss (L_{proj}) for the mesh in RoboSMPLX helps to achieve lower PVE and MPJPE errors. It facilitates the learning of more precise body model and camera parameters to improve the alignment between the rendered 3D model and 2D image. Notably, metrics such as PVE and MPJPE errors is calculated after root alignment and may not sufficiently reflect the quality of mesh projection onto the image space. To offer a more precise analysis, we evaluate the discrepancies between the projected 2D vertices of the ground-truth and projected meshes. More quantitative and qualitative comparisons can be found in Appendix F.

6 Conclusion

In this paper, we introduce a new framework RoboSMPLX to advance the field of whole-body pose and shape estimation. It enhances the whole-body pipeline by learning more precise localization for part crops while ensuring that part subnetworks are robust enough to handle suboptimal part crops and produce reliable outputs. It achieves this goal with three innovations: accurate subject localization by explicitly learning both sparse and dense predictions of the subject, robust feature extraction with supervised contrastive learning, and accurate pixel alignment of outputs with differentiable rendering. Nevertheless, it is important to acknowledge that there are instances in which our framework exhibits limitations, such as (1) inaccurate beta estimation due to out-of-distribution data (children), (2) challenges posed by severe object-occlusion, (3) difficulties arising from person-person occlusion, and (4) the potential for prediction errors in multi-person scenarios, as exemplified by the cases detailed in Appendix G. These challenges represent important avenues for future refinement of our approach.

There are several potential avenues for future research. First, the current approach does not deliberately select negative samples during training. Future work could explore if hard mining by intentionally selecting similar poses in a batch could enhance learning. Second, the careful selection of augmentations is essential. While augmentations that modify the global orientation, such as flipping and rotation, have proven detrimental and are not employed, the effects of individual augmentations and their combinations are not examined. Future research could explore the potential for automatically determining the optimal selection of augmentations to achieve improved performance. Additionally, simplifying the complex framework without sacrificing performance is a beneficial direction for future work. Lastly, considering that videos are a prevalent input format, the integration of video-based estimation can contribute to bolstering model robustness can enhance model robustness, alleviate depth ambiguity, and improve temporal consistency.

Acknowledgements

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-PhD-2023-08-049T). This study is also supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOE-T2EP20221-0012), NTU NAP, and under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s). We sincerely thank the anonymous reviewers for their valuable comments on this paper.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3686–3693, 2014. ISSN 10636919. doi: 10.1109/CVPR.2014.471.

[2] Yutong Bai, Angtian Wang, Adam Kortylewski, and Alan Yuille. CoKe: Contrastive Learning for Robust Keypoint Detection. *Proceedings - 2023 IEEE Winter Conference on Applications of Computer Vision, WACV 2023*, pp. 65–74, 2023. doi: 10.1109/WACV56688.2023.00015.

[3] Adnane Boukhayma, Rodrigo De Bem, and Philip H.S. Torr. 3D hand shape and pose from images in the wild. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:10835–10844, 2019. ISSN 10636919. doi: 10.1109/CVPR.2019.01110.

[4] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2Mesh: Graph Convolutional Network for 3D Human Pose and Mesh Recovery from a 2D Human Pose. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12352 LNCS:769–787, 2020. ISSN 16113349. doi: 10.1007/978-3-030-58571-6_45.

[5] Hongsuk Choi, Hyeongjin Nam, Taeryung Lee, Gyeongsik Moon, and Kyoung Mu Lee. Rethinking Self-Supervised Visual Representation Learning in Pre-training for 3D Human Pose and Shape Estimation. *International Conference on Learning Representations (ICLR)*, pp. 1–18, 2023. URL http://arxiv.org/abs/2303.05370.

[6] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision (ECCV)*, 2020. URL https://expose.is.tue.mpg.de.

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

[8] Radek Danecek, Michael Black, and Timo Bolkart. EMOCA: Emotion Driven Monocular Face Capture and Animation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022-June:20279–20290, 2022. ISSN 10636919. doi: 10.1109/CVPR52688.2022.01967.

[9] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3D face reconstruction with weakly-supervised learning: From single image to image set. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2019-June:285–295, 2019. ISSN 21607516. doi: 10.1109/CVPRW.2019.00038.

[10] Sai Kumar Dwivedi, Nikos Athanasiou, Muhammed Kocabas, and Michael J. Black. Learning to Regress Bodies from Images using Differentiable Semantic Rendering. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11230–11239, 2021. ISBN 9781665428125. doi: 10.1109/iccv48922.2021.01106.

[11] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision (3DV)*, 2021.

[12] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. In *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, volume 40, 2021. URL https://doi.org/10.1145/3450626.3459936.

[13] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6M. *Ieee Transactions on Pattern Analysis and Machine intelligence*, pp. 1, 2014. ISSN 01628828. URL http://109.101.234.42/documente/publications/1-82.pdf.

[14] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-Body Human Pose Estimation in the Wild. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12354 LNCS:196–214, 2020. ISSN 16113349. doi: 10.1007/978-3-030-58545-7_12.

[15] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 8320–8329, 2018. ISSN 10636919. doi: 10.1109/CVPR.2018.00868.

[16] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar Fine-Tuning for 3D Human Model Fitting Towards In-the-Wild 3D Human Pose Estimation. *Proceedings - 2021 International Conference on 3D Vision, 3DV 2021*, pp. 42–52, 2021. doi: 10.1109/3DV53792.2021.00015.

[17] 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 Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 7122–7131, 2018. ISBN 9781538664209. doi: 10.1109/CVPR.2018.00744.

[18] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43 (12):4217–4228, 2021. ISSN 19393539. doi: 10.1109/TPAMI.2020.2970919.

[19] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.

[20] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Proc. International Conference on Computer Vision (ICCV)*, pp. 11127–11137, October 2021.

[21] Muhammed Kocabas, Chun-Hao P. Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J. Black. SPEC: Seeing people in the wild with an estimated camera. In *Proc. International Conference on Computer Vision (ICCV)*, pp. 11035–11045, October 2021.

[22] Nikos Kolotouros, Georgios Pavlakos, Michael Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-Octob:2252–2261, 2019. ISSN 15505499. doi: 10.1109/ICCV.2019.00234.

[23] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:4496–4505, 2019. ISSN 10636919. doi: 10.1109/CVPR.2019.00463.

[24] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic Modeling for Human Mesh Recovery. In *International Conference on Computer Vision (ICCV)*, pp. 11585–11594, 2021. ISBN 9781665428125. doi: 10.1109/iccv48922.2021.01140.

[25] Lingdong Kong, Youquan Liu, Xin Li, Runnan Chen, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Robo3d: Towards robust and reliable 3d perception against corruptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19994–20006, 2023.

[26] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. URL https://doi.org/10.1145/3130800.3130813.

[27] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. *arXiv preprint arXiv:2303.16160*, 2023.

[28] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-End Human Pose and Mesh Reconstruction with Transformers. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1954–1963, 2021. ISSN 10636919. doi: 10.1109/CVPR46437.2021.00199.

[29] 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. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8693 LNCS(PART 5):740–755, 2014. ISSN 16113349. doi: 10.1007/978-3-319-10602-1_48.

[30] Qihao Liu, Adam Kortylewski, and Alan L Yuille. Poseexaminer: Automated testing of out-of-distribution robustness in human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 672–681, 2023.

[31] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

[32] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2019. ISSN 19493045. doi: 10.1109/TAFFC.2017.2740923.

[33] Gyeongsik Moon and Kyoung Mu Lee. I2L-MeshNet: Image-to-Lixel Prediction Network for Accurate 3D Human Pose and Mesh Estimation from a Single RGB Image. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12352 LNCS:752–768, 2020. ISSN 16113349. doi: 10.1007/978-3-030-58571-6_44.

[34] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *European Conference on Computer Vision (ECCV)*, 2020.

[35] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2022.

[36] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. NeuralAnnot: Neural Annotator for 3D Human Mesh Training Sets. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2022-June:2298–2306, 2022. ISSN 21607516. doi: 10.1109/CVPRW56347.2022.00256.

[37] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. *Proceedings - 2018 International Conference on 3D Vision, 3DV 2018*, pp. 484–494, 2018. doi: 10.1109/3DV.2018.00062.

[38] Hui En Pang, Zhongang Cai, Lei Yang, Tianwei Zhang, and Ziwei Liu. Benchmarking and analyzing 3d human pose and shape estimation beyond algorithms. In *NeurIPS*, 2022.

[39] 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 *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

[40] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to Estimate 3D Human Pose and Shape from a Single Color Image. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 459–468, 2018. ISSN 10636919. doi: 10.1109/CVPR.2018.00055.

[41] 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 *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, pp. 10967–10977, 2019. ISBN 9781728132938. doi: 10.1109/CVPR.2019.01123.

[42] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11488–11499, 2021.

[43] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), November 2017.

[44] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. FrankMocap: A Monocular 3D Whole-Body Pose Estimation System via Regression and Integration. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2021-Octob, pp. 1749–1759, 2021. ISBN 9781665401913. doi: 10.1109/ICCVW54120.2021.00201.

[45] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[46] Adrian Spurr, Aneesh Dahiya, Xi Wang, Xucong Zhang, and Otmar Hilliges. Self-Supervised 3D Hand Pose Estimation from monocular RGB via Contrastive Learning. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 11210–11219, 2021. ISSN 15505499. doi: 10.1109/ICCV48922.2021.01104.

[47] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11210 LNCS:536–553, 2018. ISSN 16113349. doi: 10.1007/978-3-030-01231-1_33.

[48] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. Recovering 3D Human Mesh from Monocular Images: A Survey. *arXiv preprint arXiv:2203.01923*, pp. 1–20, 2022. URL http://arxiv.org/abs/2203.01923.

[49] Hsiao Yu Fish Tung, Hsiao Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. *Advances in Neural Information Processing Systems*, 2017-December(Nips):5237–5247, 2017. ISSN 10495258.

[50] 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. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11214 LNCS:614–631, 2018. ISSN 16113349. doi: 10.1007/978-3-030-01249-6_37.

[51] Jiahang Wang, Sheng Jin, Wentao Liu, Weizhong Liu, Chen Qian, and Ping Luo. When human pose estimation meets robustness: Adversarial algorithms and benchmarks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 11850–11859, 2021. ISSN 10636919. doi: 10.1109/CVPR46437.2021.01168.

[52] Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 9319–9328, 2020. ISSN 10636919. doi: 10.1109/CVPR42600.2020.00934.

[53] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:10957–10966, 2019. ISSN 10636919. doi: 10.1109/CVPR.2019.01122.

[54] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM GHUML: Generative 3D human shape and articulated pose models. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 6183–6192, 2020. ISSN 10636919. doi: 10.1109/CVPR42600.2020.00622.

[55] Yuanlu Xu, Song Chun Zhu, and Tony Tung. DenseRaC: Joint 3D pose and shape estimation by dense render-and-compare. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-Octob:7759–7769, 2019. ISSN 15505499. doi: 10.1109/ICCV.2019.00785.

[56] Lei Yang, Qingqiu Huang, Huaiyi Huang, Linning Xu, and Dahua Lin. Learn to propagate reliably on noisy affinity graphs. In *European Conference on Computer Vision*, pp. 447–464. Springer, 2020.

[57] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly Supervised 3D Human Pose and Shape Reconstruction with Normalizing Flows. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12351 LNCS:465–481, 2020. ISSN 16113349. doi: 10.1007/978-3-030-58539-6_28.

[58] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[59] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-Occluded Human Shape and Pose Estimation from a Single Color Image. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 7374–7383, 2020. ISSN 10636919. doi: 10.1109/CVPR42600.2020.00740.

[60] Yumeng Zhang, Li Chen, Yufeng Liu, Xiaoyan Guo, Wen Zheng, and Junhai Yong. Improving robustness for pose estimation via stable heatmap regression. *Neurocomputing*, 492:322–342, 2022. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2022.04.046. URL https://www.sciencedirect.com/science/article/pii/S0925231222004131.

[61] Yuxiao Zhou, Marc Habermann, Ikhsanul Habibie, Ayush Tewari, Christian Theobalt, and Feng Xu. Monocular Real-time Full Body Capture with Inter-part Correlations. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, number 61822111 in 61822111, pp. 4809–4820, 2021. ISBN 9781665445092. doi: 10.1109/CVPR46437.2021.00478.

[62] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max J. Argus, and Thomas Brox. FreiHAND: A dataset for markerless capture of hand pose and shape from single rgb images. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-October: 813–822, 2019. ISSN 15505499. doi: 10.1109/ICCV.2019.00090.

[63] Christian Zimmermann, Max Argus, and Thomas Brox. Contrastive Representation Learning for Hand Shape Estimation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13024 LNCS:250–264, 2021. ISSN 16113349. doi: 10.1007/978-3-030-92659-5_16.

Overview

This supplementary material presents more details and additional results not included in the main paper due to the page limitation. The list of items included are:

- Description of augmentation settings for robustness benchmarking in Section A
- More experiment setup and details in Section B
- Comparison with PyMAF-X in Section C
- Analysis of the subpar performance on AGORA test set in Section D
- Ablation of different modules on the Body subnetwork in Section E
- Quantitative and qualitative and comparisons for pixel alignment in Section F
- Examples of failure cases in Section G
- Analysis of embedding similarity in Section H
- Discussion on pose (rotation) versus keypoint representation in Section I
- Extra comparisons against SOTA body networks in Section J
- Training and inference time in Section K
- Accuracy of derived part bounding boxes in Section L
- Qualitative comparisons of `RoboSMPLX`'s Hand, Face and Body subnetworks under augmentations in Section M
- Quantitative and qualitative comparisons of `RoboSMPLX`'s wholebody model in Section N

A Augmentation Settings for Robustness Benchmarking

In the selection of augmentations, we opted for a set of ten commonly encountered augmentations that could be benchmarked in a controlled setting. We also ensure that the selected values for manipulation fall within a realistic range. We used the following augmentations:

1. Vertical translation: We shifted the image by factors relative to the image size. For instance, a +0.1 shift corresponds to a 10% upward movement, while a -0.1 shift represents a 10% downward movement. Our boundaries were set at ± 0.3 to ensure that majority of the subject remains visible within the image frame.
2. Horizontal translation: We manipulated the image by factors relative to the image size. A shift of +0.1 denotes a 10% move to the right, while -0.1 indicates a 10% shift to the left. We imposed a ± 0.3 limit to keep the majority of the subject within the image.
3. Scale: We adjusted the person's crop using factors relative to the bounding box size. For example, a factor of +0.1 leads to a 10% size reduction, resulting in a tighter crop, while a -0.1 factor enlarges the crop size by 10%. A ± 0.5 boundary was set to maintain visibility of the majority of the person within the image.
4. Low Resolution: The resolution of the cropped image was modified by factors related to the image size. A 2.0 factor signifies that the image was downsampled to half its original size before being upsampled back, reducing the resolution by a factor of 2.0.
5. Rotation: The image was manipulated by various rotations up to degrees of ± 60.
6. Hue: The image hue was altered by converting the image to HSV format, cyclically shifting intensities in the hue channel (H), and converting back to the original image mode. Hue adjustments were limited to ± 0.5.
7. Sharpness: Sharpness was controlled by introducing an enhancement factor. A factor of -1.0 leads to a blurred image, while +1.0 results in a sharpened image, with 0.0 leaving the image unaltered. This effect is achieved by blending the source image with the degraded mean image.
8. Grayness: The degree of grayness was adjusted by introducing an enhancement factor. A factor of -1.0 results in a completely grayed image, while +1.0 leads to a whitened image, with 0.0 leaving the image unaltered. This effect is achieved by blending the source image with its gray counterpart. The limit was set to ± 0.5, as the subject becomes unidentifiable at extremes of ± 1.0.

9. Contrast: This was controlled by introducing an enhancement factor. A factor of -1.0 leads to a completely grayed image, while +1.0 results in a whitened image, with 0.0 leaving the image unaltered. This effect is achieved by blending the source image with the degraded mean image. The limit was set to ± 0.5, as the subject becomes unidentifiable at extremes of ± 1.0.
10. Brightness: The brightness of the image was adjusted by introducing an enhancement factor. A factor of -1.0 results in a black image, while +1.0 leads to a white image, with 0.0 leaving the image unaltered. This effect is achieved by blending the source image with the degraded black image. The limit was set to ± 0.5, as the subject becomes unidentifiable at extremes of ± 1.0.

B More Experiment Setup

This section includes extra description of each submodule and implementation details.

Body subnetwork. The body image is downsampled from the original image to reduce the computational cost, resulting in $I_b \in R^{3\times 256\times 256}$. The Body subnetwork outputs 3D body joint rotations $\theta_b \in R^{21\times 3}$, global orientation $\theta_{bg} \in R^3$, shape parameters $\beta_b \in R^{10}$, camera parameters $\pi_b \in R^3$, and whole-body joints $K \in R^{137\times 3}$. Hand and face bounding boxes are then derived from the face and hand keypoints. Width and height are determined from the x-y range of the keypoints, and the center is the aggregated mean of the keypoints. High resolution crops are used for hand and face inputs following ExPose and PIXIE. In line with ExPose [6] and PIXIE [11], hand and face input images are obtained from high resolution crops to utilize the information available from the original image instead of the downsampled image.

Hand subnetwork. After obtaining the cropped hand images $I_h \in R^{3\times 256\times 256}$, the left hand images are flipped to match the orientation of the right hands before being input to the Hand subnetwork. After predicting the 3D finger rotations $\theta_h \in R^{15\times 3}$, the outputs of the flipped left hands are reverted to their original orientation. The 3D finger rotations of the left and right hands are denoted as θ_{rh} and θ_{lh} respectively. When training the full version on hand datasets, we also output the global orientation $\theta_{hg} \in R^3$, shape $\beta_h \in R^{10}$ and camera $\pi_h \in R^3$. However, these branches are discarded during whole-body estimation and training.

Face subnetwork. This subnetwork generates the 3D jaw rotation $\theta_f \in R^3$ and expression $\psi_f \in R^{10}$ from the cropped face image $I_f \in R^{3\times 256\times 256}$. When training the full version on face datasets, additional outputs include the global orientation $\theta_{fg} \in R^3$, shape $\beta_f \in R^{50}$, expression $\psi_f \in R^{50}$ and camera $\pi_f \in R^3$. These branches are also discarded during whole-body estimation and training.

Implementation details. The training and evaluation of our model builds upon the MMHuman3D framework [7]. For model initialization, we pre-train the ResNet backbone on the MSCOCO 2D whole-body human pose dataset. During training, we use the Adam optimizer with a mini-batch size of 32 and apply data augmentations, e.g., scaling, rotation, random horizontal flip, and color jittering. The initial learning rate is set to $10\mathrm{e}{-4}$, decayed by a factor of 10 at the later epoch. We use the SMPL, MANO, FLAME and SMPL-X body models for the training of body, hand, face and wholebody respectively. Further details will be provided in our code.

C Comparison with PyMAF-X

Below we provide detailed discussions and comparisons with PyMAF-X [58].

1. Acquisition of part bounding boxes: PyMAF-X relies on an off-the-shelf whole-body pose estimation model (OpenPifpaf) to obtain whole body 2D keypoints of the person in the image, from which part crops are derived. During the EHF evaluation, PyMAF-X employs hand and face bounding boxes derived from OpenPose keypoints. In contrast, our method and other works (ExPose [6], PIXIE [11], Hand4Whole [35] and OS-X [27].) encompass a self-integrated module designed to extract hand and face bounding boxes directly from the image.
2. Operational efficiency: Openpifpaf imposes extra computation during inference, making PyMAF-X less efficient than our method. Please refer to Section K in Appendix.
3. Network architecture: Due to the diverse backbone and dataset combinations utilized, it is challenging for us to make whole-body network comparisons. In Table 1, we focus

on contrasting RoboSMPLX's Hand subnetwork with PyMAF's Hand subnetwork. Both networks are trained and evaluated on the same backbone and dataset, FreiHAND. In this context, our method surpasses PyMAF.

4. Performance: On the EHF metrics, our performance lags behind PyMAF-X. This could potentially arise from variations in the training datasets employed. While the training pipeline of the body network for PyMAF-X has been disclosed, the training specifics for hands and face and the methodology to integrate hand, face, and body module PyMAF-X, remains undisclosed. We intend to replicate with similar training datasets in the future.

D Analysis of performance on AGORA test set

Figure 13 visualise samples with significant errors during training. AGORA contains extensive person-to-person occlusion, frequently leading to substantial overlap between the target individual (marked with red vertices) and another person. In cases that experienced large errors, the model often incorrectly identified the target individual as the person situated in the forefront (model predictions marked with green vertices), thereby introducing instability throughout the training process due to the model's challenge in accurately discerning the intended subject.

We also added qualitative comparisons of `RoboSMPLX` under varying scales and alignments as shown in Figures 14. We demonstrate that `RoboSMPLX` produces better pixel alignment of the body, and more accurate hand and face predictions where the target person has been accurately identified.

Figure 13: **Visualisation of samples with high errors at train time.** Red vertices indicates the target person while green vertices are the model's predictions.

Figure 14: **Visualisation of Expose [41], PIXIE [11], Hand4Whole [35], OS-X [27] and `RoboSMPLX` under different scales and alignment on AGORA validation set.**

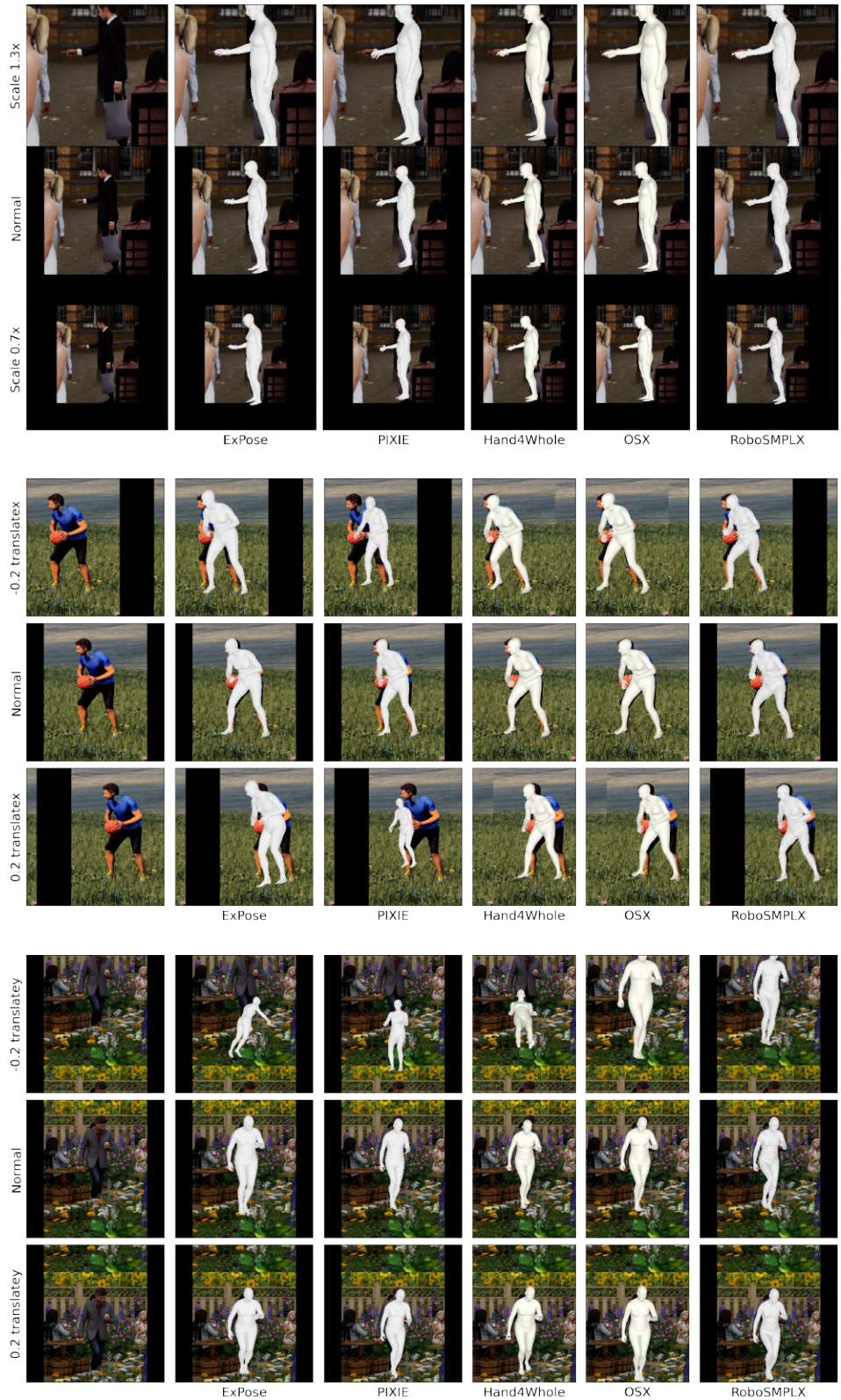

Figure 14: **Visualisation of Expose [41], PIXIE [11], Hand4Whole [35], OS-X [27] and `RoboSMPLX` under different scales and alignment on AGORA validation set (cont.).**

E Ablation on Body Subnetwork

Table 12 shows the ablation of different modules on the Body subnetwork. The conclusions derived from the Hand ablation study (Table 11) extends to the Body subnetwork as well.

Table 12: **Ablation of different modules on Body subnetwork. Results are trained on EFT-COCO and tested on 3DPW test set.**

	loss	representation	PA-	MPJPE
Baseline (HMR)	-	-	60.8	96.2
LF (all)	-	-	56.7	105.7
LF (all), L_{con}	L1	pose	55.9	90.9
LF (all), L_{con}	MSE	pose	58.5	93.9
LF (all), L_{con}	SmoothL1	pose	56.6	92.5
LF (all), L_{con}	L1	pose(rot6d)	58.9	95.0
LF (all), L_{con}	L1	pose + go	76.8	118.9
LF (all), L_{con}, +ve	L1	keypoints	**55.4**	**90.56**

F Qualitative and quantitative comparisons for pixel alignment

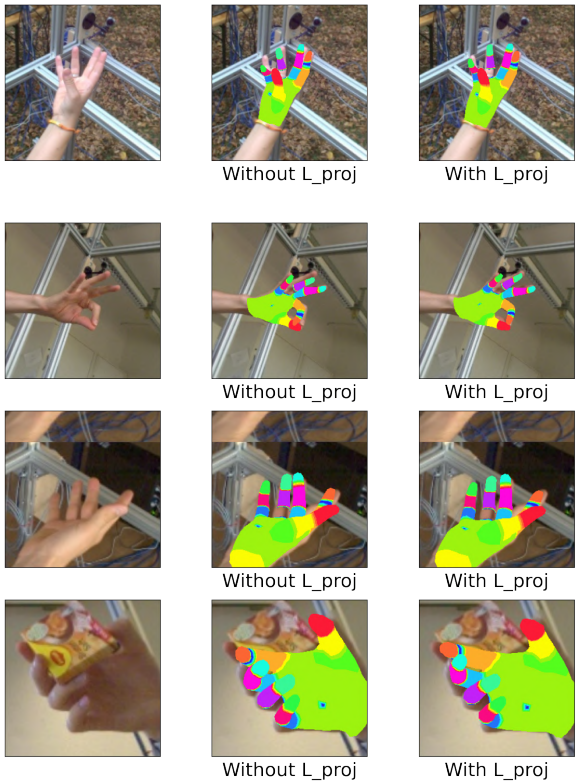

Figure 15: **(C) Visualisation from training with and without L_{proj}.**

Prevailing metrics such as Per Vertex Error (PVE) and Mean Per Joint Position Error (MPJPE) do not incorporate alignment measurement in their evaluation. Before these metrics are computed, the mesh undergoes root alignment, but this process does not necessarily reflect the level of alignment accuracy when the mesh is reprojected back into the image space.

Moreover, for pose and shape estimation methods, the absence of ground-truth camera parameters implies that there is no direct supervision for these parameters. Camera parameters are, instead, often weakly supervised through the supervision of projected keypoints (derived from regressed joints of the mesh and predicted camera parameters) and the ground-truth 2D joints by ensuring their alignment. This only provides a sparse supervision. To enhance better learning of camera, pose and shape parameters, pixel alignment strategy is introduced, which ensures denser supervision.

Presently, there's an absence of a metric tailored to gauge the degree of pixel alignment of a mesh in this context. We included qualitative examples of training with and without L_{proj}, and demonstrate that the projection of vertices results in better pixel alignment (Figure 15).

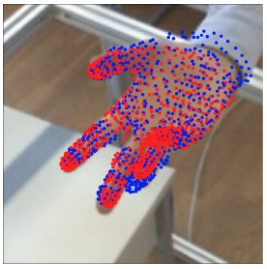

Figure 16: **Projected Vertex Errors is measured as distance between projected ground-truth (red) and predicted (blue) vertices in image space.**

Method	Projected Vertex Errors ↓
HMR (no PA)	11.796
HMR + PA (vertex)	11.211
HMR + PA (part-seg)	**10.298**

Table 13: **Results of Projected Vertex Errors under different Part Alignment (PA)**

To provide quantitative analysis, we measure errors between the projected 2D vertices of ground-truth and projected meshes (Figure 16). From Table 13, it is evident that omitting the pixel alignment module leads to suboptimal outcomes. In contrast, our pixel alignment strategy, leveraging rendered segmentation maps, showcases better performance than using vertex loss as supervision.

G Failure cases

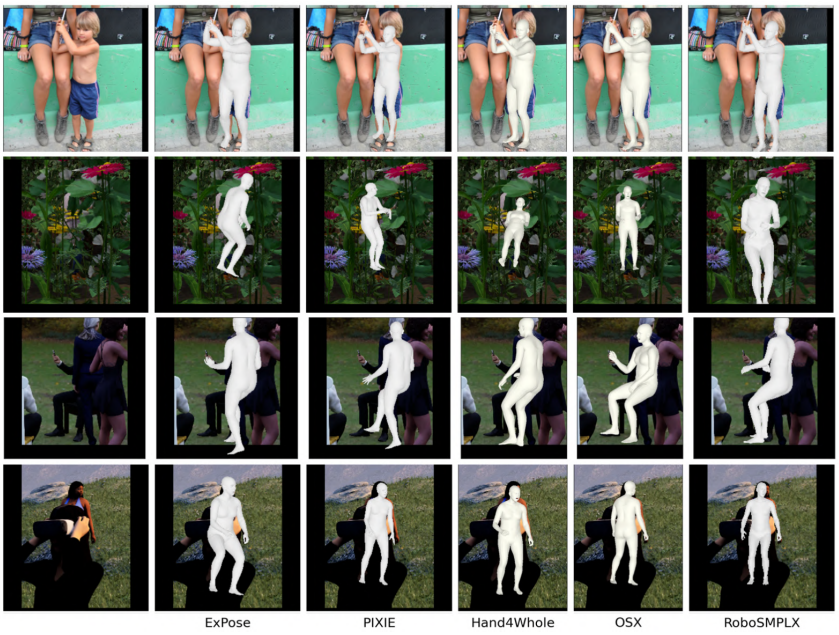

Figure 17: **Examples of failure cases. (1) Inaccurate beta estimation due to out-of-distribution data (children) (2) Severe object-occlusion (3) Person-person occlusion (4) Prediction for wrong person in multi-person scenarios.**

H Embedding similarity

Our use of the contrastive module is motivated by the need to constrain/maintain the same pose feature for different augmentations, to avoid domain shift caused by strong augmentation alone. The experiments show that the use of strong augmentation alone for training can lead to performance deterioration, while combining it with the contrastive loss consistently results in minimal errors (Table 11).

Table 14: **Ablation of CFE module on Hand Subnetwork. (This excludes Localization and Pixel Alignment Module). Results are trained and evaluated on FreiHAND.**

Method	PA-MPJPE ↓	MPJPE ↓	PA-PVE ↓	PVE ↓	Pose embedding distance ↓
Model 0: HMR	8.06	16.78	7.85	16.71	0.132
Model 1: HMR + Strongaug	8.47	17.01	8.11	16.17	0.138
Model 2: HMR + Strongaug + CL	**7.79**	**15.68**	**7.41**	**15.27**	0.101

To illustrate this further, we delved into a visualization of the pose similarity for augmented samples. The findings reveal that augmented samples are perceived as dissimilar in both Model 0 and Model 1 (Table 14). Yet, when examining Model 2, a marked increase in embedding similarity is evident, underscoring the advantage of the contrastive approach.

I Discussion of pose versus keypoint representation

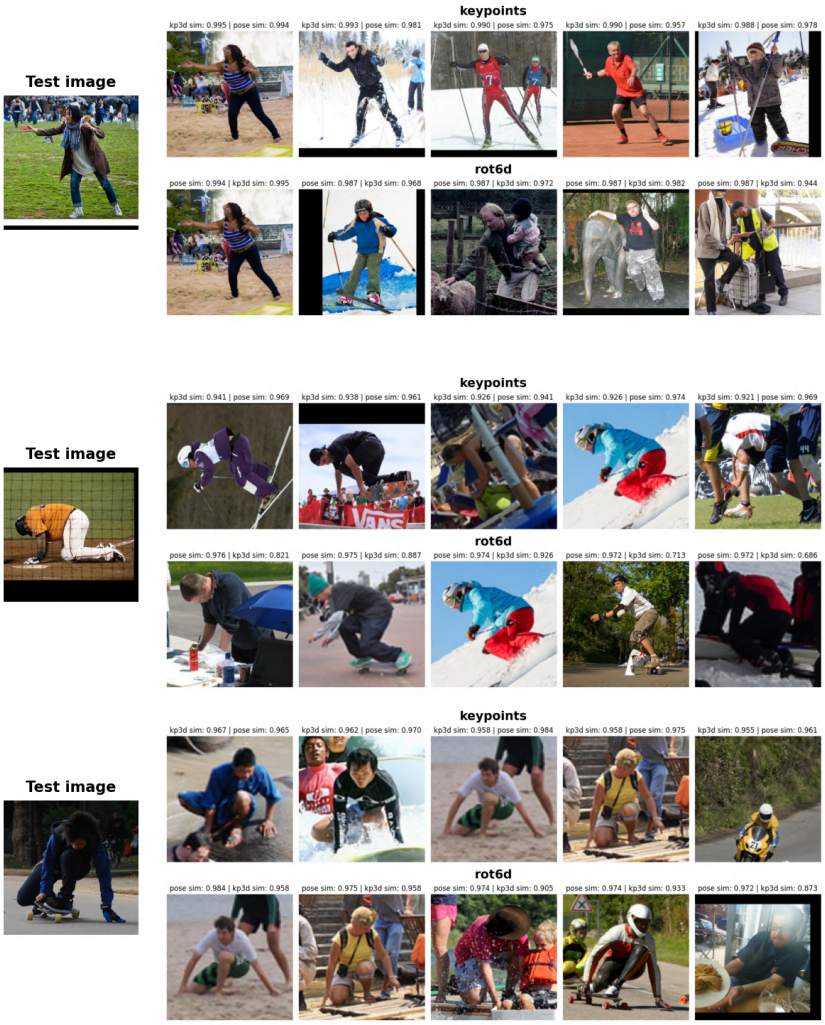

Figure 18: **Comparison of keypoints and pose representations.**

Figure 18 compares two distinct methods for image retrieval: one based on pose similarity (rot6d representation) and the other based on keypoint similarity. Samples with high keypoint similarity tends to have comparatively high pose similarity. On the contrary, similar pose representation might have considerably lower joint representation. This could occur due to the accumulation of minor discrepancies in joint rotations which, over time, may result in significant disparities in the keypoints.

This results in instances where the overall pose of the retrieved sample is very high to the query sample, but the keypoints may not coincide as accurately. Meanwhile, using keypoint representation would result in samples that have demonstrate improved alignment with the query image, presenting a more accurate correspondence.

This could explain why using regressed keypoints as representation have better performance (Table 9). Clustering based on keypoint similarity is more effective than pose similarity, as pose representation might be susceptible to minor shifts in joint rotations.

J Extra comparisons against SOTA body networks

There are many factors affecting training,including, but not limited to, the choice of backbone, datasets employed, and specific protocols executed during evaluation. Specifically, with regards to 3DPW, various protocols—ranging from fine-tuning (3DPW Protocol 1), collective training, to omission during training (3DPW Protocol 2)—have a large influence on 3DPW results in the evaluation process.

In Table 15, we outperform HybrIK when using the same backbone (HRNet-W48) and not fine-tuning on 3DPW (3DPW Protocol 2). Notably, CLIFF incorporated 3DPW within its training datasets. Given that our approach and that of both HybrIK and CLIFF do not utilize identical dataset combinations, a direct comparison becomes inherently challenging.

Table 15: Evaluation of HybrIK, CLIFF and our network on 3DPW. Our results are also available in Table 2.

Method	Backbone	F-T on 3DPW	PA-MPJPE (3DPW)	MPJPE (3DPW)
HybrIK	HRNet-W48	No	48.6	88.0
HybrIK	HRNet-W48	Yes	41.8	71.3
CLIFF	Res-50	Trained with 3DPW	45.7	72.0
CLIFF	HRNet-W48	Trained with 3DPW	43.0	69.0
Ours	Resnet-50	No	49.8	80.8
Ours	HRNet-W48	No	48.5	80.1

We have provided qualitative comparisons of body-only methods under different scale and alignment in Figure 21. Below, we provide quantitative evaluations of our method with HMR, SPIN and PARE (Table 16). Our method is able to achieve better performance under different scales and alignment.

Table 16: Evaluated on 3DPW (PA-MPJPE/MPJPE) under different scales and alignment. * denote the same dataset combination

	Normal	Transx +0.2x	Transx -0.2x	Transy +0.2y	Transy -0.2y	Scale 1.3x	Scale 0.7x
HMR	67.53/112.34	77.31/141.70	77.06/ 138.51	86.57/ 151.15	77.26/148.33	68.46/ 117.1	75.38/ 124.79
SPIN	57.54/94.11	70.14/122.56	68.67/ 120.04	73.08/ 111.33	70.64/133.2	61.08/ 103.60	61.63/ 99.6
PARE (HR32) *	**49.3/81.8**	74.9/139.2	77.1/ 141.7	59.1/92.3	64.2/ 109.7	54.7/86.9	**50.5/ 83.9**
Ours (R50) *	49.8/80.8	**67.2/117.2**	**67.72/111.5**	**56.4/90.0**	**62.8/105.6**	**50.2/84.6**	50.8/ 82.4

K Training and inference time

Our model was trained utilizing a cluster of 8xTesla V100-SXM2-32GB GPUs. Specific to the training duration, the hand models required approximately one day, whereas the body and face models necessitated two days. The joint training process was completed within a day.

We measure the model size, computation complexity and inference time for different models including ours, as shown in Table 17. Although our framework has sophisticated design, it has comparable inference speed as others, validating its efficacy.

L Quantitative evaluation of predicted bounding box accuracy

To assess the precision of predicted part bounding boxes on the EHF test set, we utilized Intersection over Union (IoU) as our evaluation metric (see Figure 19). Our method achieved the highest IoU

Table 17: These results are tested on RTX3090. FLOP refers to the total number of floating point operations required for a single forward pass. The higher the FLOPs, the slower the model and hence low throughput. Inference Time is obtained by averaging across 100 runs.

	Total parameters (M)	GFLOPs	Inference time (s)
ExPose	26.06	21.04	0.1330 ± 0.0050
PIXIE	109.67	24.23	0.1670 ± 0.0065
Hand4Whole	77.84	17.98	0.0709 ± 0.0022
OSX	422.52	83.77	0.1998 ± 0.0028
PyMAF-X (gt H/F bbox)	205.93	33.41	0.2194 ± 0.0027
PyMAF-X + OpenPipaf	205.93 + 115.0	33.41 + 120.52	0.2727 ± 0.0136
RoboSMPLX	120.68	29.66	0.2008 ± 0.0220

scores, as demonstrated in Table 18. It is important to note that in the OSX implementation, the hand and face features are cropped from the body features rather than directly from the image.

Figure 19: **Calculation for the Face, LHand and RHand IoU scores for ground-truth (green) and predicted (red) part bounding boxes.**

Method	Face IoU	LHand IoU	RHand IoU
ExPose	0.61	0.23	0.31
PIXIE	0.66	0.34	0.36
Hand4Whole	0.75	0.41	0.45
OSX	0.70	0.38	0.41
RoboSMPLX	0.86	0.52	0.55

Table 18: **Results for IoU of the predicted part bounding boxes on the EHF test set.**

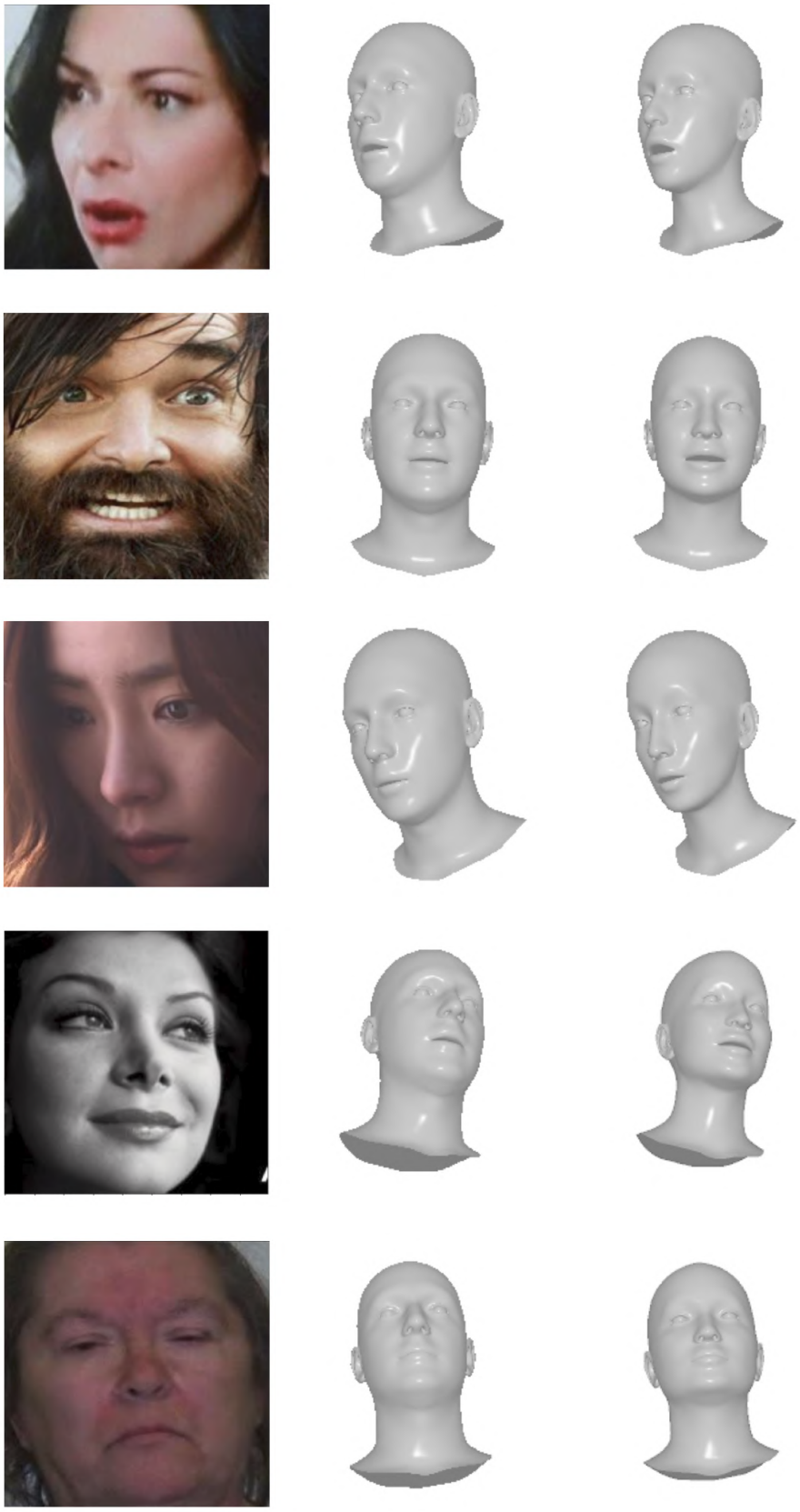

Figure 8: **Inference on AffectNet validation images using Expose [6] and `RoboSMPLX`'s Face subnetwork.**

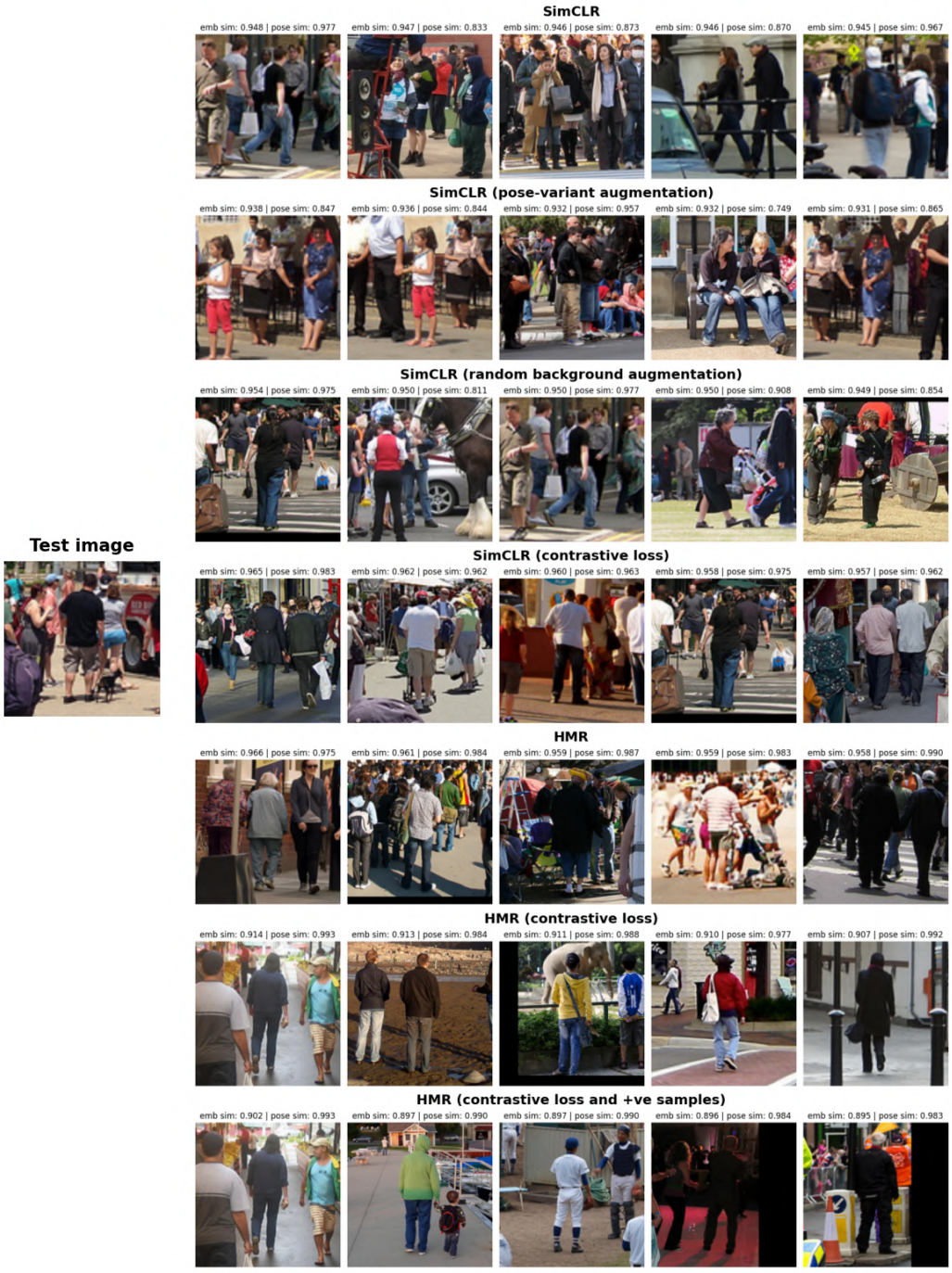

Figure 9: **Left: Query image from the EFT-COCO-Test set, Right: Retrieved image from the EFT-COCO-Train set ordered in descending embedding similarity.**

Figure 10: **Left: Query image from the EFT-COCO-Test set, Right: Retrieved image from the EFT-COCO-Train set ordered in descending embedding similarity.**

Figure 11: **Left: Query image from the EFT-COCO-Test set, Right: Retrieved image from the EFT-COCO-Train set ordered in descending embedding similarity.**

Figure 12: **Left: Query image from the EFT-COCO-Test set, Right: Retrieved image from the EFT-COCO-Train set ordered in descending embedding similarity.**

M Qualitative comparisons for different models under augmentation

We show qualitative comparisons of RoboSMPLX's Hand (Figure 20), Face (Figure 21) and Body (Figure 21) subnetwork to existing models under different positional augmentations.

In general, RoboSMPLX s' subnetworks demonstrate better pixel alignment and are less sensitive to changes in scale and alignment.

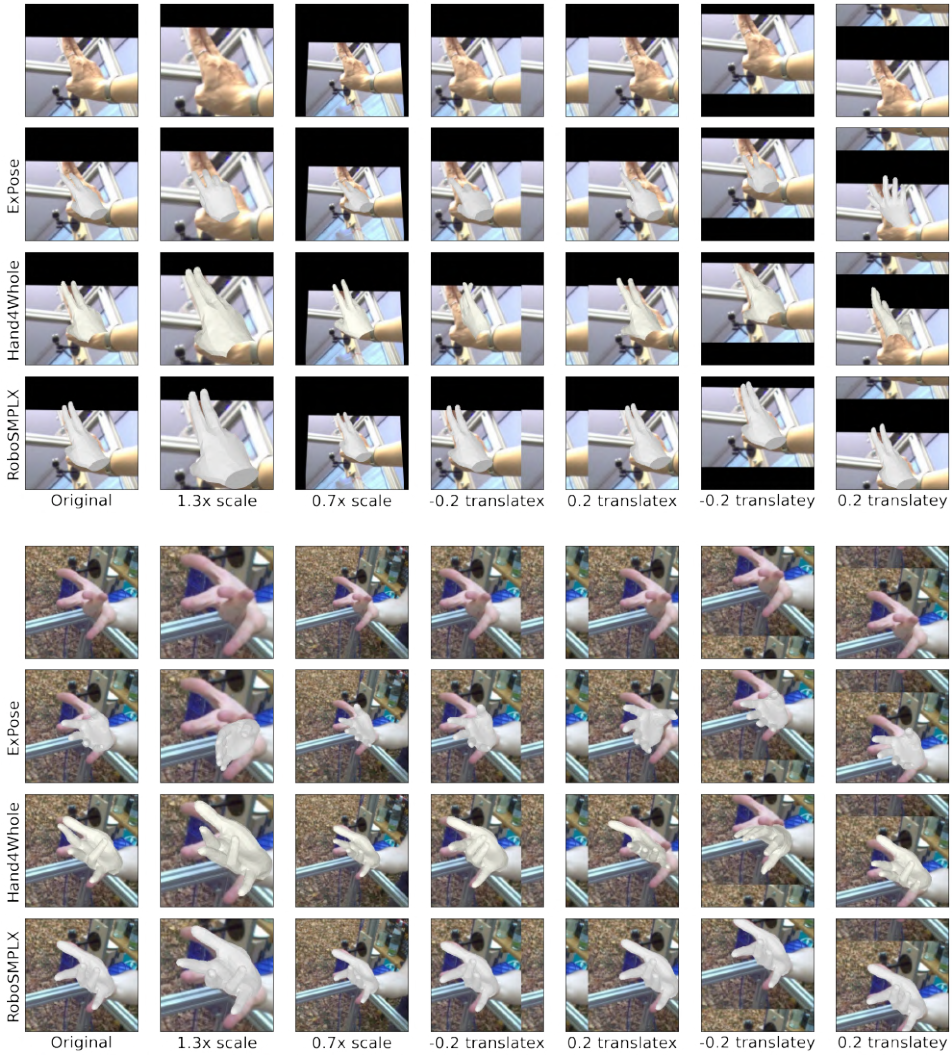

Figure 20: **Comparison of ExPose [6], Hand4Whole [35] and RoboSMPLX's Hand subnetwork under various augmentations on FreiHAND test set.**

Figure 20: **Comparison of of ExPose [6], Hand4Whole [35] and `RoboSMPLX`'s Hand subnetwork under various augmentations on FreiHAND test set (cont.)**

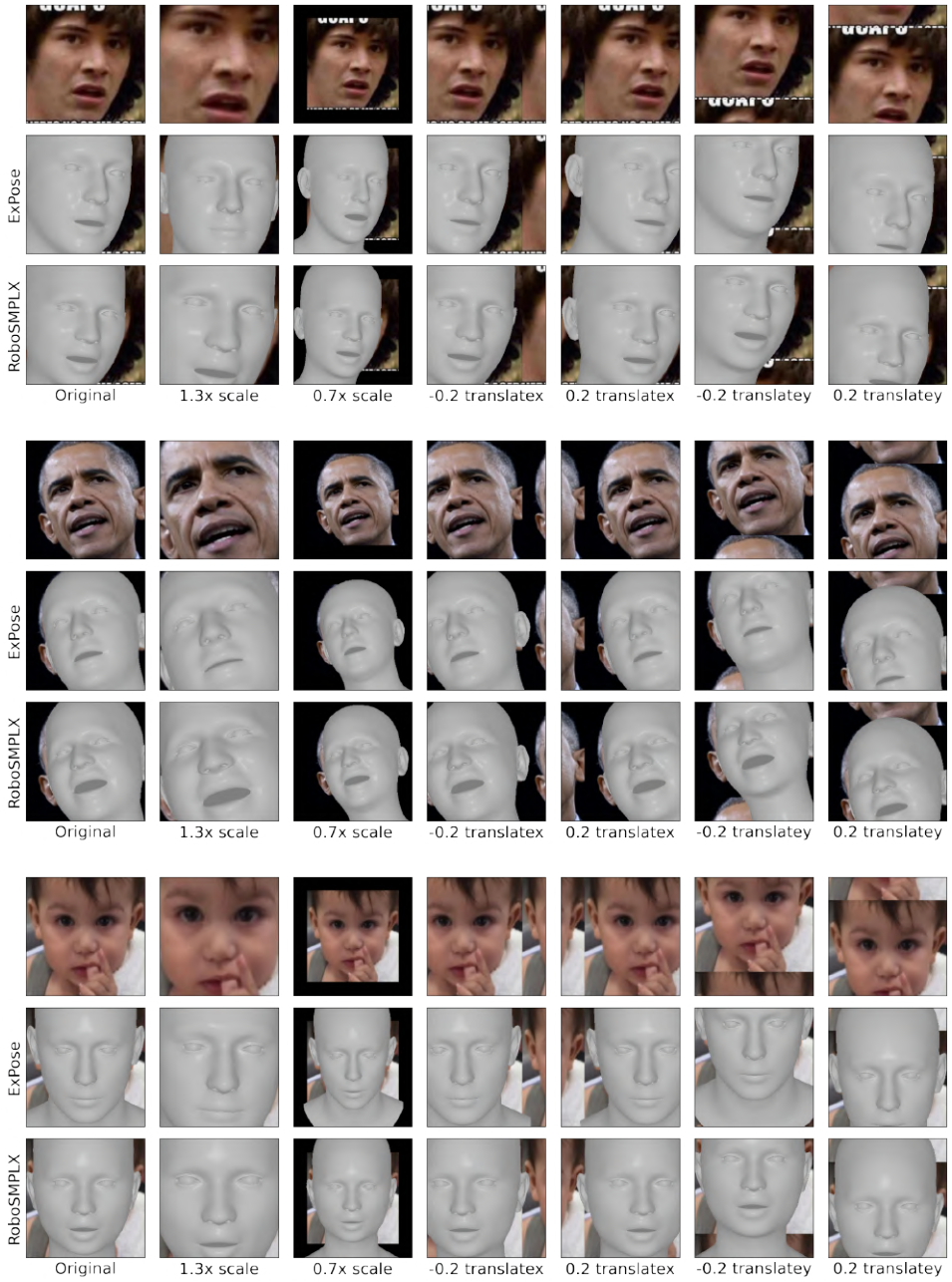

Figure 21: **Comparison of ExPose [6] and `RoboSMPLX`'s Face subnetwork under various augmentations on AffectNet val set.**

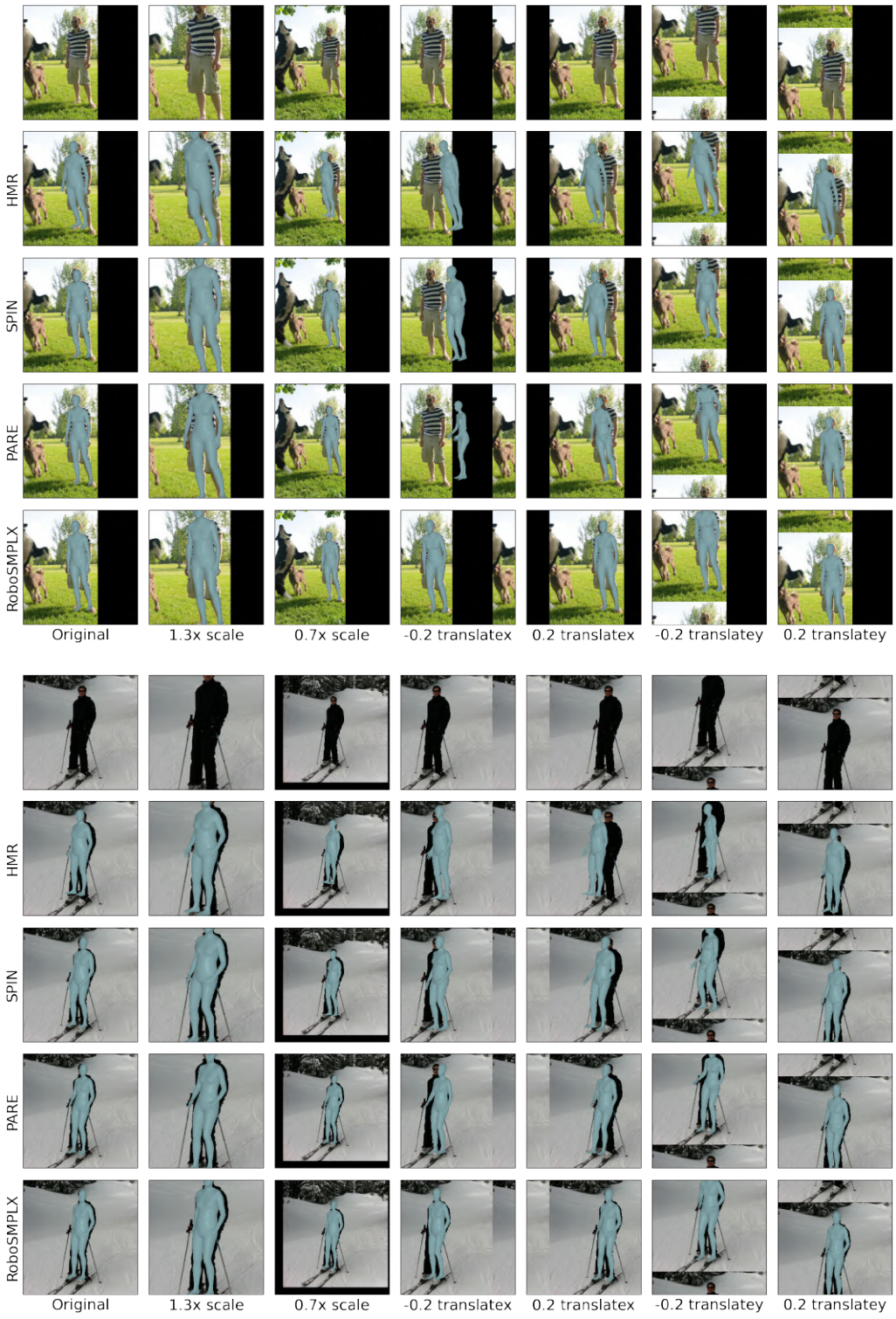

Figure 21: **Comparison of HMR [17], SPIN [22], PARE[20] and RoboSMPLX's Body subnetwork under various augmentations on COCO validation set.**

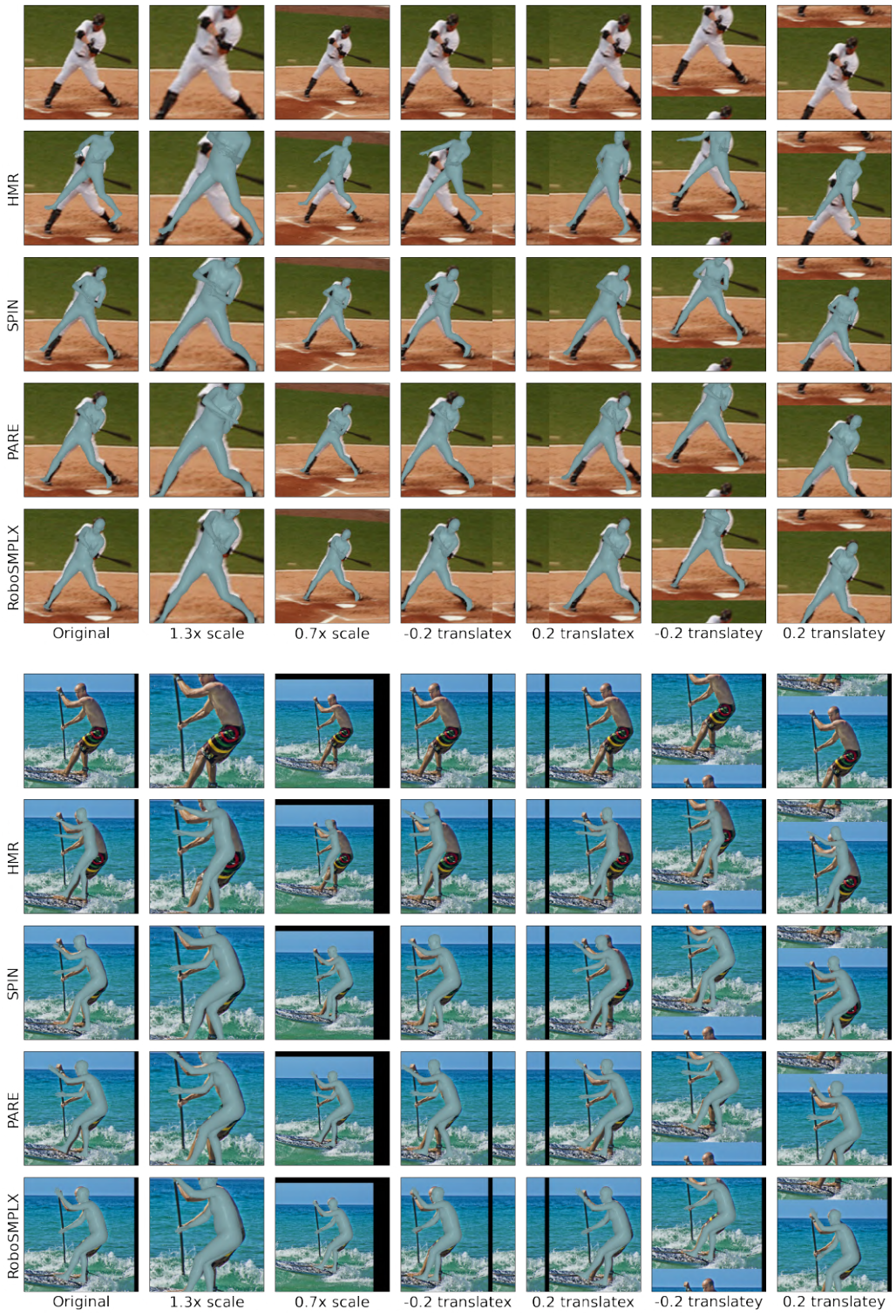

Figure 21: **Comparison of HMR [17], SPIN [22], PARE[20] and RoboSMPLX's Body subnetwork under various augmentations on COCO validation set (cont.)**

N Quantitative and qualitative comparisons for wholebody models

We provide quantitative comparisons of wholebody models under different augmentations on EHF test set in Figures 22 and 23. We also added qualitative comparisons under different scale and alignment on EHF test set in Figures 24 to 26. We demonstrate that RoboSMPLX produces better pixel alignment of the body, and more accurate hand and face predictions. In addition, we inference on in-the-wild examples on COCO-validation set in Figures 27 and 28.

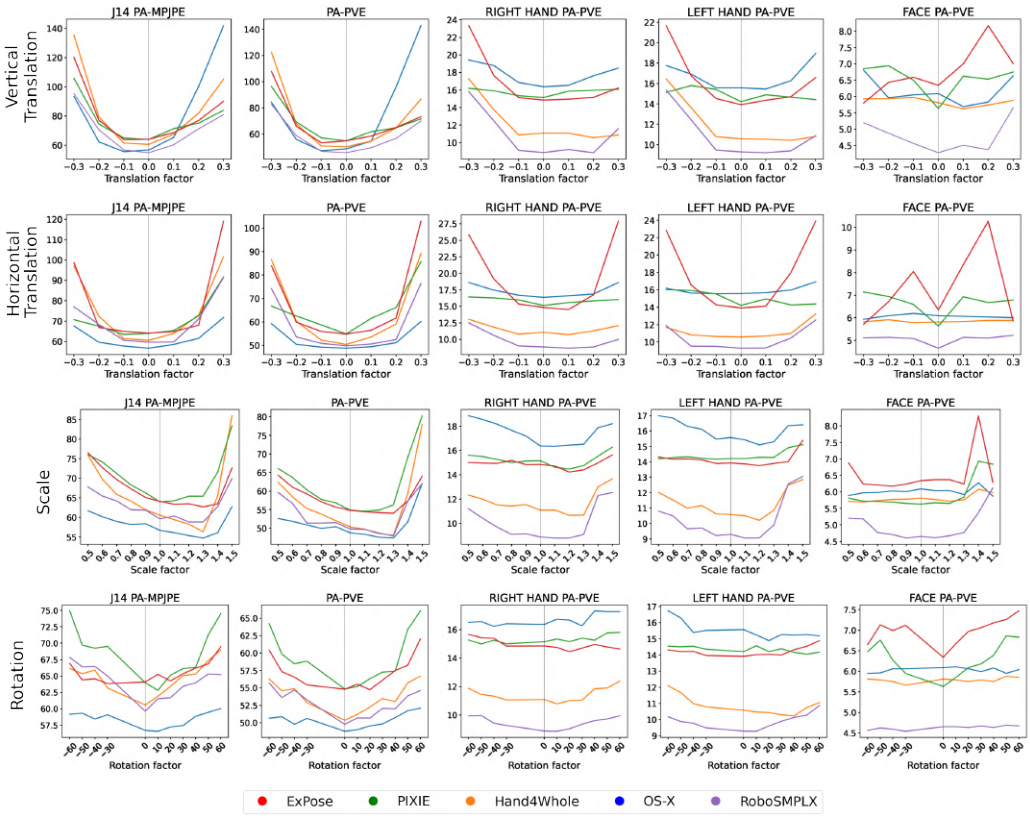

Figure 22: **Wholebody errors under different amounts of augmentation on EHF test set. The gray line indicates baseline performance without augmentation.**

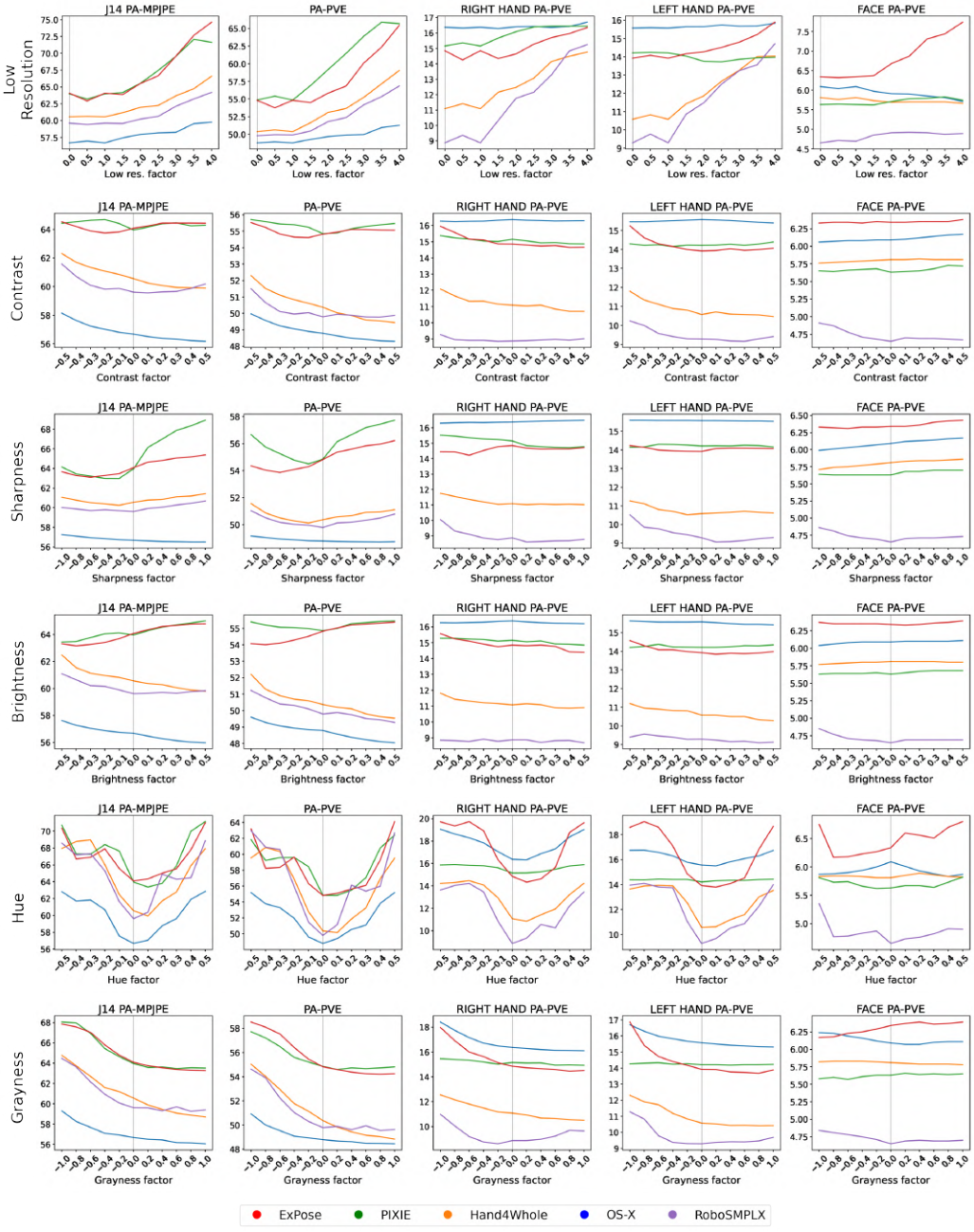

Figure 23: **Wholebody errors under different amounts of augmentation on EHF test set (cont.) The gray line indicates baseline performance without augmentation.**

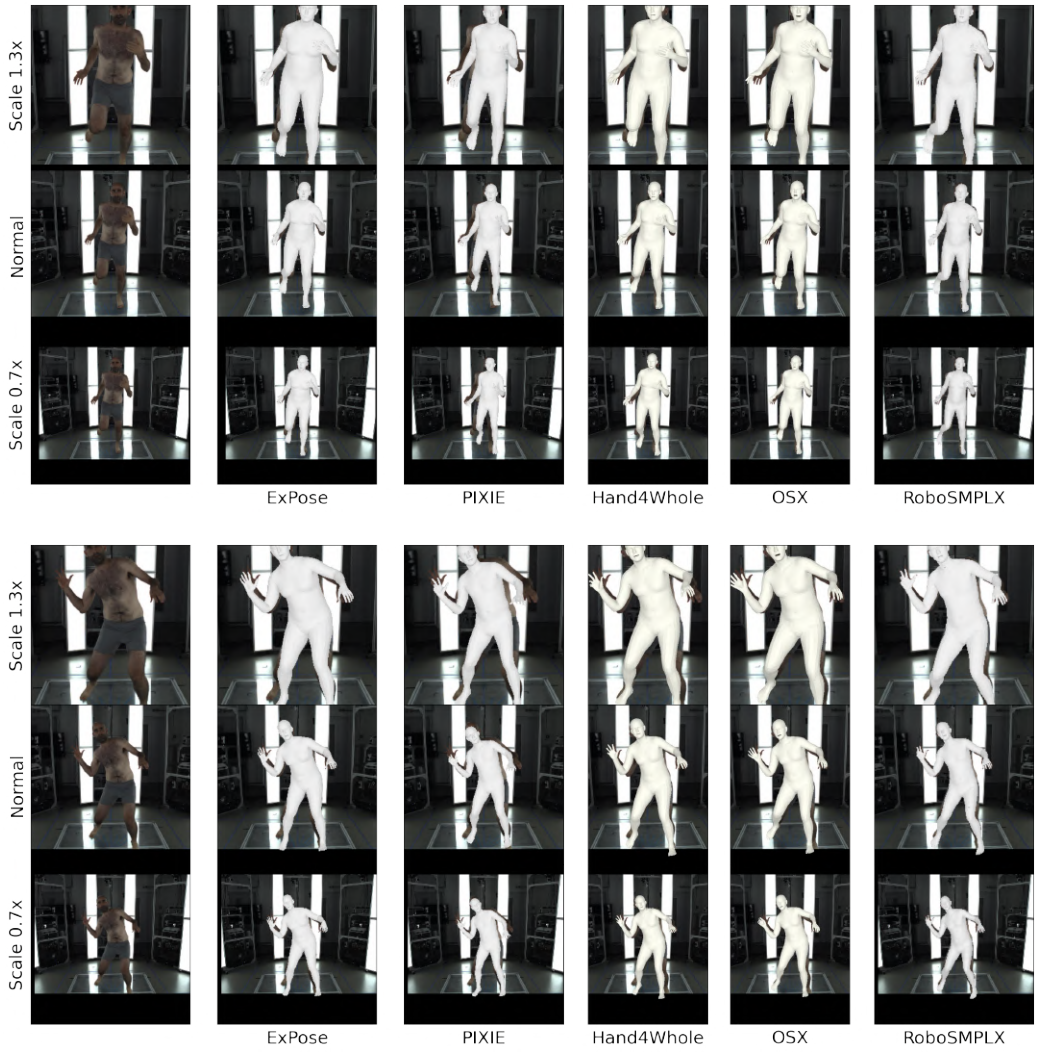

Figure 24: **Visualisation of Expose [41], PIXIE [11], Hand4Whole [35], OS-X [27] and `RoboSMPLX` under different scales on EHF test set.**

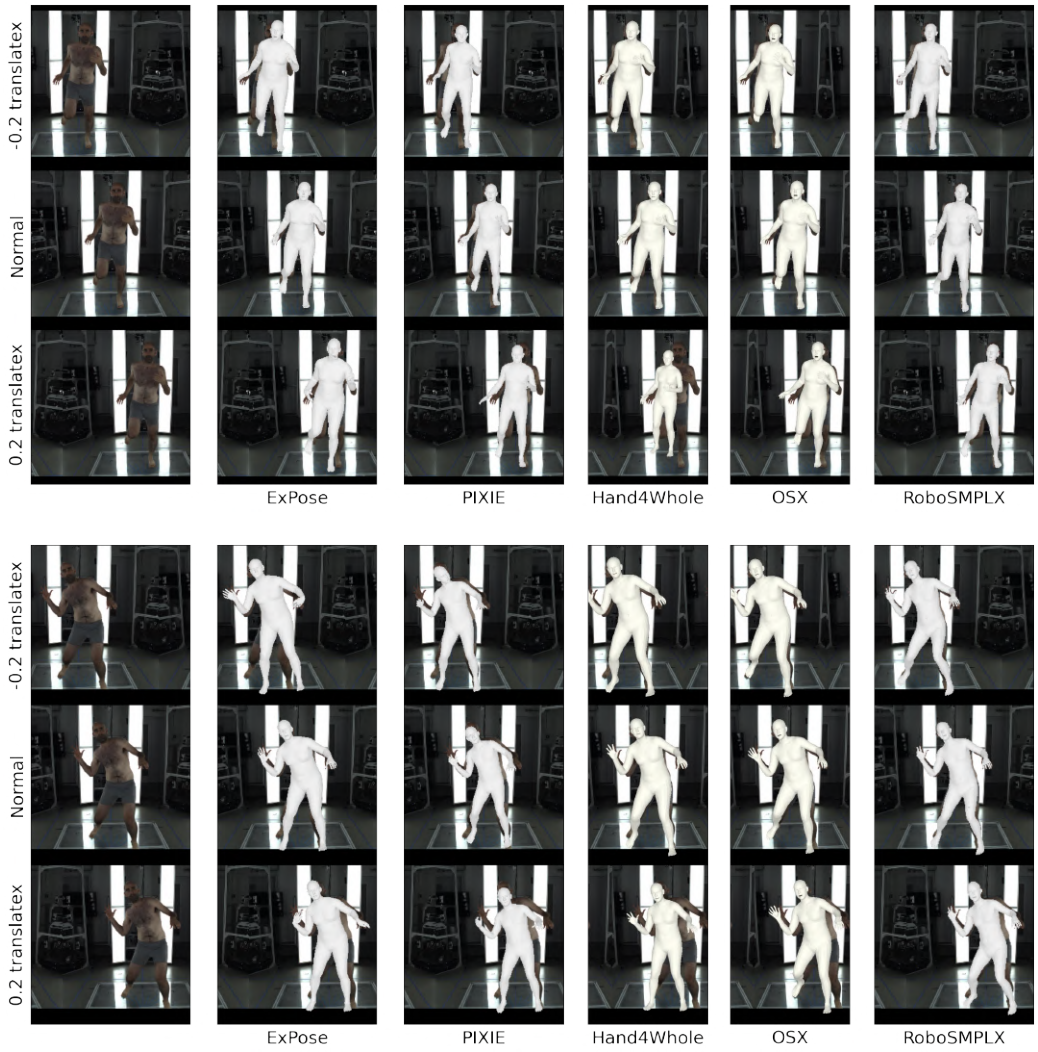

Figure 25: **Visualisation of Expose [41], PIXIE [11], Hand4Whole [35], OS-X [27] and `RoboSMPLX` under different levels of horizontal translation on EHF test set.**

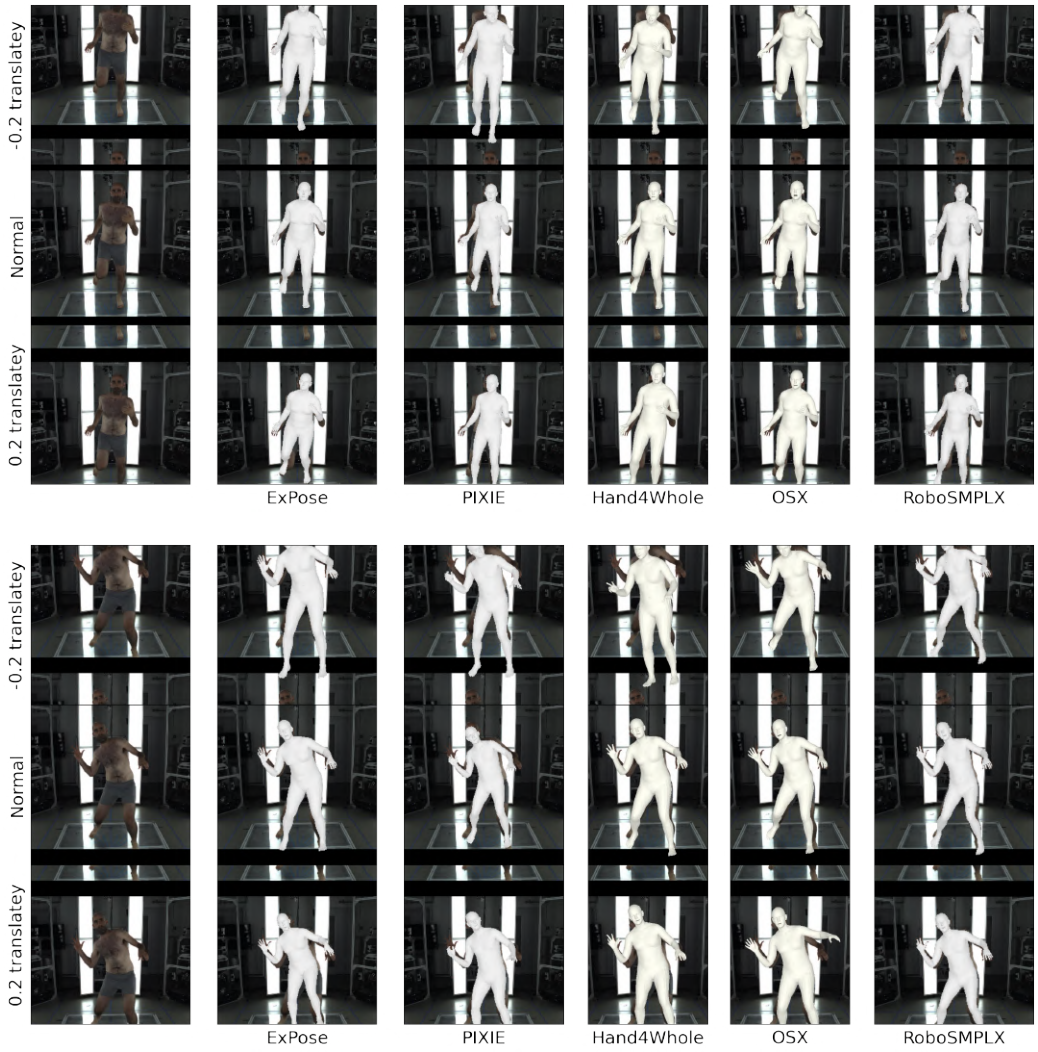

Figure 26: **Visualisation of Expose [41], PIXIE [11], Hand4Whole [35], OS-X [27] and `RoboSMPLX` under different levels of vertical translation on EHF test set.**

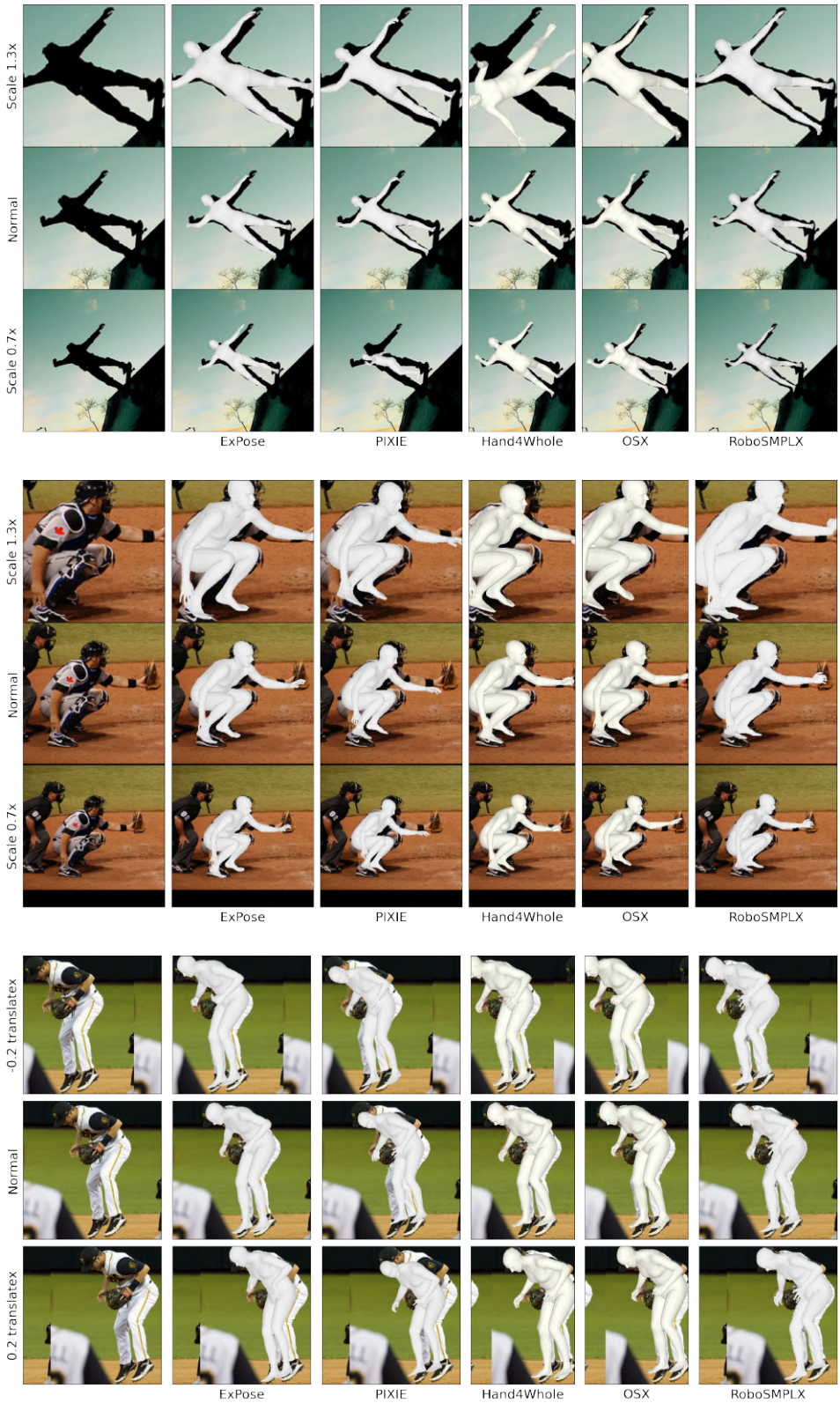

Figure 27: **Visualisation of Expose [41], PIXIE [11], Hand4Whole [35], OS-X [27] and `RoboSMPLX` under different scales and alignment on COCO validation set.**

Figure 28: **Visualisation of Expose [41], PIXIE [11], Hand4Whole [35], OS-X [27] and `RoboSMPLX` under different scales and alignment on COCO validation set.**

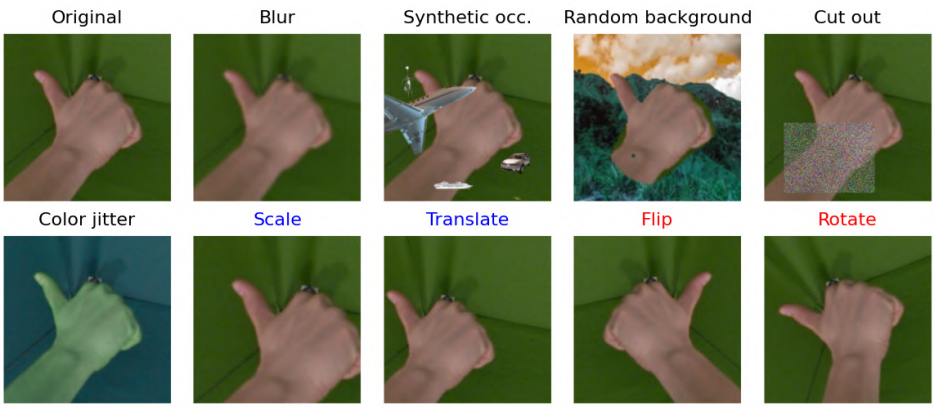

Figure 29: **Augmentations for Hand sub-networks. Blue and red labels represent location-variant and pose-variant augmentations respectively.**