Supplementary Material: Self-supervised Human Mesh Recovery with Cross-Representation Alignment

Xuan Gong[1,2], Meng Zheng[2], Benjamin Planche[2], Srikrishna Karanam[2], Terrence Chen[2], David Doermann[1], and Ziyan Wu[2]

[1] University at Buffalo, Buffalo NY, USA
xuangong@buffalo.edu, doermann@buffalo.edu
[2] United Imaging Intelligence, Cambridge MA, USA
{first.last}@uii-ai.com

This document provides material supplementing the main manuscript. Section A details the processing and augmentation of synthetic data during training. Section B contains ablation studies w.r.t. cross-representation alignment with noise perturbation, as well as different model structures for off-the-shelf detectors. Section C provides more qualitative results with comparisons of the one-representation method, baseline method of two representations, and the proposed method with cross-representation alignment.

A Training Data Synthesis and Augmentation

We generate paired data on the fly to train the regression network from intermediate representations to human mesh. The overall process can be divided into sampling, projection and rendering, and augmentation on intermediate representations. We describe the details below and provide the values of hyper-parameters in Table S1.

Sampling for mesh synthesis. We sample the pose parameters $\boldsymbol{\theta}$ from MoCap priors of UP-3D [2], 3DPW [3], and Human3.6M [1] training set. We sample the shape parameters $\boldsymbol{\beta}$ from independent normal distribution $\beta_n \sim \mathcal{N}(\mu_n, \sigma_n^2)(n = 1, \ldots, 10)$. We forward the pose and shape parameters into the SMPL model and obtain the vertices of a human mesh. We extract $N_{\mathrm{J}} = 17$ COCO 3D joints from the vertices.

Projection and rendering. To obtain the intermediate representations (*i.e.*, 2D joints and IUV map) from the human mesh, we fix the focal length as intrinsic camera parameters and sample the camera rotation and translation as extrinsic parameters for perspective projection. With these camera parameters, we can project the 3D joint to 2D joint representations. Besides, we randomly perturb the vertices $\boldsymbol{v}$ to generalize a diverse range of human shapes. From perturbed vertices and sampled camera parameters, we render a 2D IUV map using the Pytorch3D library [4].

Augmentation of 2D representation. We detect the foreground body area on the 2D IUV map and crop around the foreground area with a bounding box for consistency between training and testing. We perform zero-padding

	Hyper-parameter	Value
Sampling	pose $\boldsymbol{\theta}$	MoCap priors
	shape $\boldsymbol{\beta}$ mean	[0.2056, 0.3356, -0.3507, 0.3561, 0.4175, 0.0309, 0.3048, 0.2361, 0.2091, 0.3121]
	shape $\boldsymbol{\beta}$ std.	$[1.25] \times 10$
Rendering	vertex perturbation mean	[0, 0, 0] m
	vertex perturbation variances	[0.01, 0.01, 0] m
	camera rotation	identity 3×3 matrix
	camera translation mean	[0,0,42] m
	camera translation variances	[0.05, 0.05, 5] m
	focal length	[5000, 5000] pixel
Augmentation 2D Representation	bbox scale range	(1.0, 1.4)
	bbox center perturbation mean	[0, 0] pixel
	bbox center perturbation variances	[5, 5] pixel
	coarse body part occlusion prob.	$[0.1] \times 6$
	fine body part occlusion prob.	$[0.05] \times 24$
	remove associated joints of occluded parts prob.	0.5
	occlusion box dimension mean	[48, 48] pixel
	occlusion box dimension variances	[24, 24] pixel
	occlusion box prob.	0.1
	2D joints L/R swap prob.	0.1
	2D joints perturbation mean	$[0\text{pixel}] \times 17$
	2D joints perturbation variances	$[8\text{pixel}] \times 17$
	remove 2D joints indices	$[7, 8, 9, 10, 13, 14, 15, 16]$
	remove 2D joints prob.	$[0.05] \times 8$

Table S1. List of hyper-parameters and values for synthetic training data generation and augmentation.

around the foreground area so that the bounding box is larger than the foreground with a scale of around 1.2. We also perturb the center of the foreground with a deviation from the center of the bounding box. Based on this bounding box augmentation strategy, we crop both IUV $\boldsymbol{M}$ map and joints heatmaps $\boldsymbol{J}$ and then resize them to the target size, *i.e.*, $H = 256$ and $W = 256$. To simulate noise and discrepancy on 2D joints and IUV prediction, we do a series of probabilistic augmentations on each of them. Similar to PartDrop in [5], we randomly occlude one of the six body parts (head, torso, left/right arm, left/right leg) in IUV maps with a coarse body part occlusion probability, randomly occlude one of the 24 body parts in IUV maps with a fine body part occlusion probability, and randomly occlude the IUV maps with a dynamically-sized rectangle. For 2D joints, we swap the left/right corresponding joints (*e.g.*, left knee and right knee) with a probability. Besides, we randomly perturb the 2D joints position with a deviation and randomly set key joints (*i.e.*, left and right elbow, wrist, knee, ankle) as invisible with a probability.

B Ablation Study

B.1 Ablations with noise perturbation

To study the efficiency of our proposed cross-representation alignment, we simulate extremely challenging conditions by adding noise on the inferred 2D joints

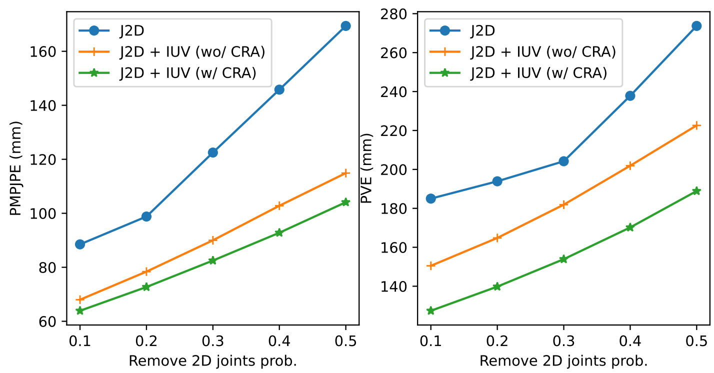

Fig. S1. Comparisons of PMPJPE and PVE when removing 2D joints with increasing probability on representations of 3DPW test images.

and IUV representations. Table 6 in the paper shows the results when adding noise on IUV and 2D joints. Figure S1 shows the comparisons with one/two representations when removing 2D joints with increasing probability. We note that using both 2D joints and IUV outperforms using 2D joint only, and the proposed cross-representation alignment can further help to improve the performance (lower PVE and PMPJPE) in the absence of 2D joints, demonstrating stronger robustness to severe noise.

Figure S2 visualizes cases when there is an occlusion in the RGB image, and the inferred IUV map fails to detect the whole body parts. Taking 2D joints and IUV representations as input, our method with cross-representation alignment (w/ CRA) can better utilize the complementarity of both representations compared with the baseline (wo/ CRA). We note that although IUV map is incomplete, the 2D joints prediction provides a sparse representation of the key points. We fully exploit the complementarity of both 2D joints and IUV map, which helps to improve the human mesh recovery result in our CRA method.

B.2 Ablations on off-the-shelf detectors

We use off-the-shelf detectors to infer 2D joints and IUV maps from RGB images for testing. For 2D joints, we use pretrained models of Keypoint-RCNN [3] and the score threshold as 0.7. For IUV, we use pretrained models of DensePose-RCNN [4]. In Table S2, we compare the results when using different backbones for 2D joints and IUV inference. It shows that different model structure designs make little difference on the 2D joints/IUV predictions and the resulting mesh

[3] https://github.com/facebookresearch/detectron2/blob/main/MODEL_ZOO.md
[4] https://github.com/facebookresearch/detectron2/blob/main/projects/DensePose/doc/DENSEPOSE_IUV.md

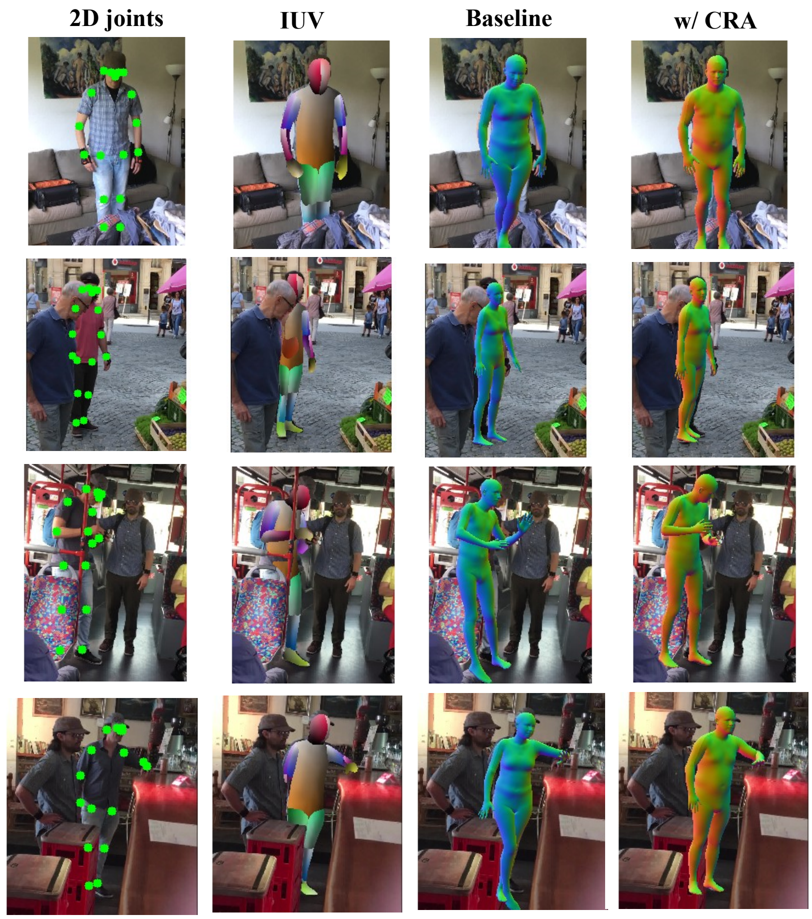

Fig. S2. Visualization of 2D joints, IUV map, baseline (wo/ CRA), and method with CRA when there is an occlusion in RGB images and resulting IUV map. Images are from the 3DPW test dataset.

Keypoint-RCNN	DensePose-RCNN	PVE↓	PMPJPE↓
ResNet50_FPN_3x	ResNet50_FPN	117.1	56.1
	ResNet101_FPN	117.5	56.2
	ResNet50_FPN_DL	117.4	56.2
	ResNet101_FPN_DL	117.4	56.3
ResNet50_FPN_1x	ResNet101_FPN_DL	118.5	57.2
ResNet50_FPN_3x		117.4	56.3
ResNet101_FPN_3x		117.2	56.5

Table S2. Comparisons of PVE and PMPJPE (both in mm) when using different model structures for Keypoint-RCNN and DensePose-RCNN to infer 2D joints and IUV on 3DPW test images. “FPN” indicates Feature Pyramid Networks, “1x” indicates training with 12 COCO epochs, “3x” indicates 3x training schedule (37 COCO epochs), and “DL” indicates DeepLabV3 head. Note no refinement is applied in this table.

recovery, demonstrating the robustness of our proposed model with respect to 2D joint detection quality. The paper reports numbers with the 2D joints/IUV inferred with ResNet50_FPN_3x for Keypoint-RCNN and ResNet101_FPN_DL for DensePose-RCNN.

C Qualitative Results

Figure S3 and Figure S4 provide more qualitative results and comparisons of mesh estimation with IUV only, with baseline method taking IUV and joints 2D as input, and the proposed method with cross-representation alignment. We can see that the mesh estimation is more likely to be biased when taking only IUV as input. When taking both IUV and joints 2D as input, the mesh estimation results improve. The additional cross-representation alignment scheme can further improve the performance with more accurate pose and shape estimation, as well as better alignment with the foreground on the RGB images.

References

1. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE transactions on pattern analysis and machine intelligence **36**(7), 1325–1339 (2013) 1
2. Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M.J., Gehler, P.V.: Unite the people: Closing the loop between 3d and 2d human representations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6050–6059 (2017) 1
3. von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: 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) 1
4. Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.Y., Johnson, J., Gkioxari, G.: Accelerating 3d deep learning with pytorch3d. arXiv:2007.08501 (2020) 1
5. Zhang, H., Cao, J., Lu, G., Ouyang, W., Sun, Z.: Learning 3d human shape and pose from dense body parts. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020) 2

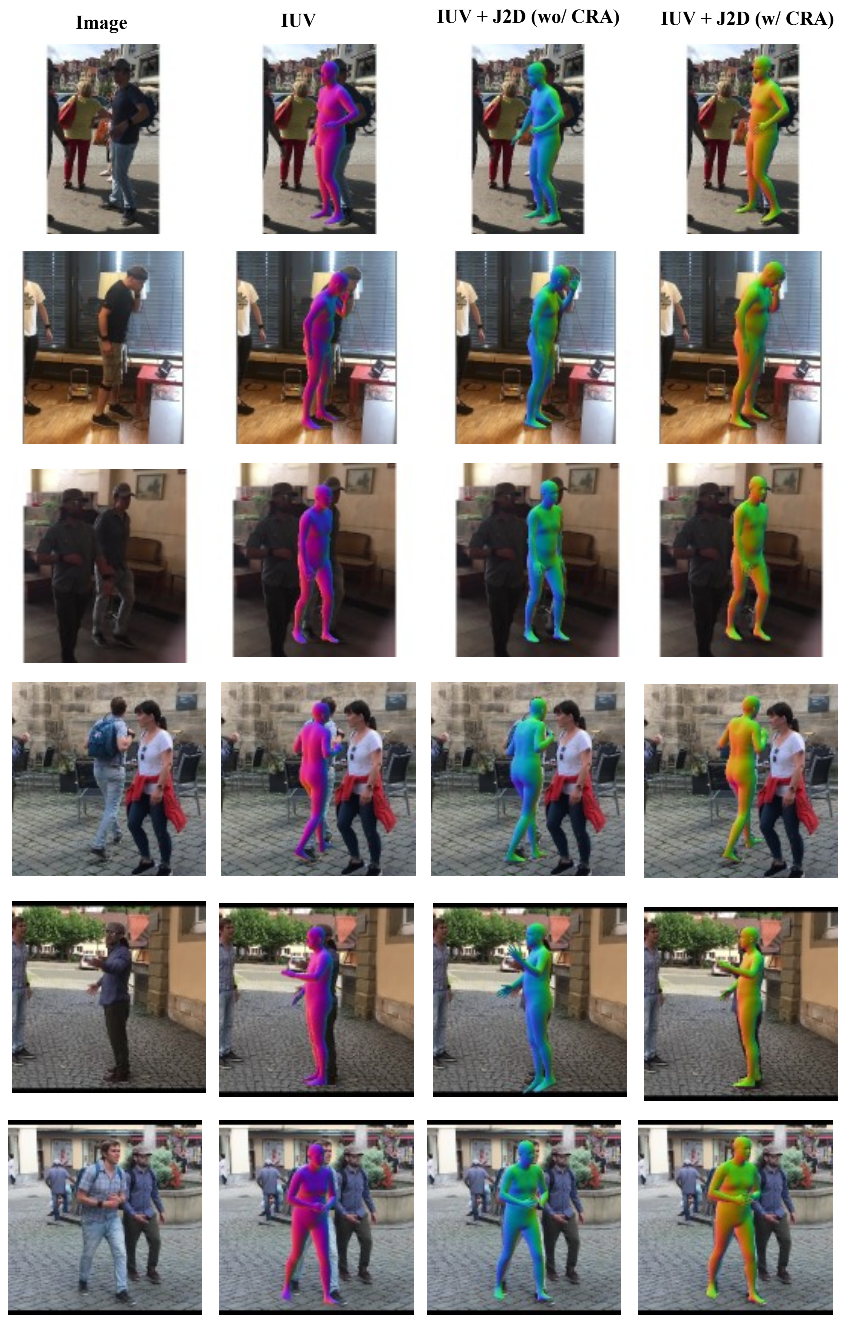

Fig. S3. Visualization of the mesh recovery results when taking only IUV as input, taking both IUV and 2D joints as input without cross-representation alignment (wo/ CRA), and taking both as input with cross-representation alignment (w/ CRA).

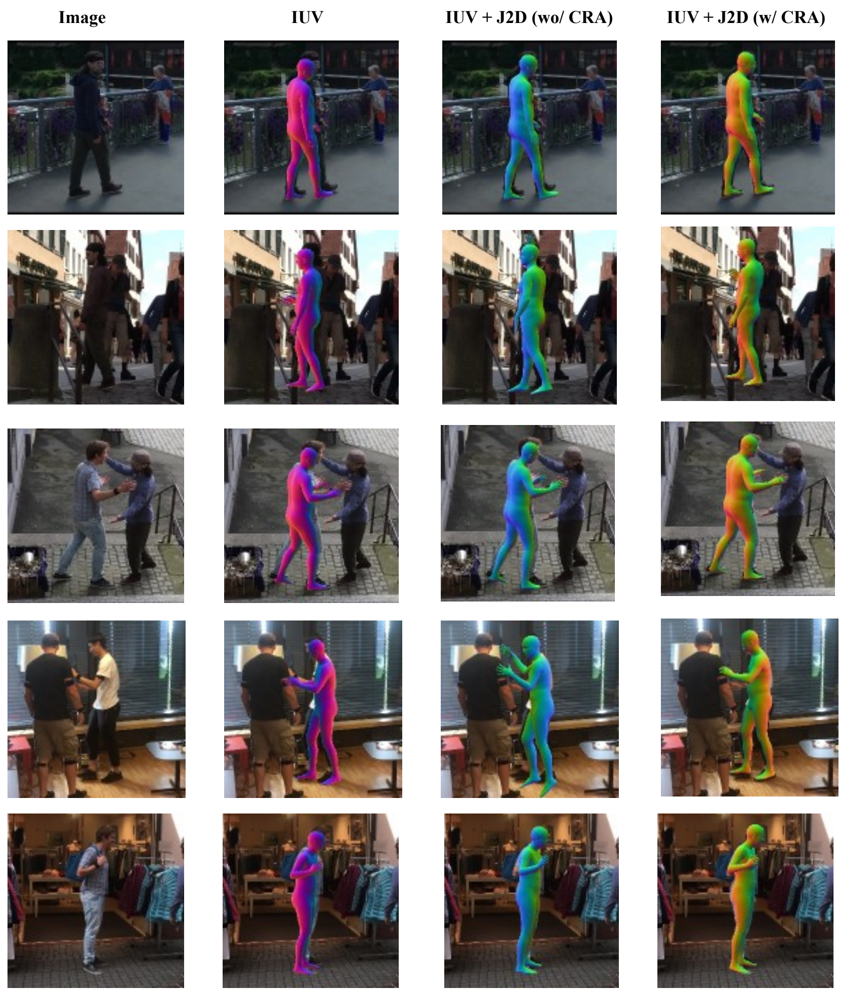

Fig. S4. Visualization of the mesh recovery results when taking only IUV as input, taking both IUV and 2D joints as input without cross-representation alignment (wo/ CRA), and taking both as input with cross-representation alignment (w/ CRA).