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

Abstract. Fully supervised human mesh recovery methods are data-hungry and have poor generalizability due to the limited availability and diversity of 3D-annotated benchmark datasets. Recent progress in self-supervised human mesh recovery has been made using synthetic-data-driven training paradigms where the model is trained from synthetic paired 2D representation (*e.g.*, 2D keypoints and segmentation masks) and 3D mesh. However, on synthetic dense correspondence maps (*i.e.*, IUV) few have been explored since the domain gap between synthetic training data and real testing data is hard to address for 2D dense representation. To alleviate this domain gap on IUV, we propose cross-representation alignment utilizing the complementary information from the robust but sparse representation (2D keypoints). Specifically, the alignment errors between initial mesh estimation and both 2D representations are forwarded into regressor and dynamically corrected in the following mesh regression. This adaptive cross-representation alignment explicitly learns from the deviations and captures complementary information: robustness from sparse representation and richness from dense representation. We conduct extensive experiments on multiple standard benchmark datasets and demonstrate competitive results, helping take a step towards reducing the annotation effort needed to produce state-of-the-art models in human mesh estimation.

Keywords: Human Mesh Recovery, Representation Alignment, Synthetic-to-Real Learning

1 Introduction

3D human analysis from images is an important task in computer vision, with a wide range of downstream applications such as healthcare [13] and computer animation [28]. We consider the problem of human mesh estimation, *i.e.*, estimating the 3D parameters of a parametric human mesh model given input data, typically RGB images. With the availability of models such as SMPL [31], there has been much recent progress in this area [2, 12, 19].

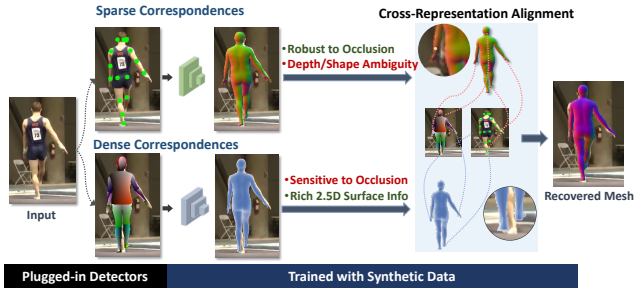

Fig. 1. Motivation: In our synthetic-data-driven pipeline, we train a model from 2D representations to 3D mesh. During test, 2D representations are inferred from off-the-shelf detectors, where sparse/dense 2D representations come with complementary advantage: 2D keypoints provide a robust but sparse representation of the skeleton, dense correspondences (IUV maps) provide rich but sensitive body information. This motivates us to explore cross-representation alignment to take advantage of both to optimize recovered human mesh.

However, obtaining good performance with these methods requires many data samples with 3D annotations. In the SMPL model, this would be the pose and shape parameters. Generating these 3D annotations is very expensive in general and prohibitive in many specific situations, such as medical settings [62]. Developing these annotations requires expensive, and custom motion capture setups and heavily customized algorithms such as MoSh [30], which are highly impractical in many scenarios, including the aforementioned medical one. This results in a situation where there are only limited datasets with 3D pose and shape annotations, further resulting in models that tend to perform well in narrow scenarios while generalizing poorly to out of distribution data [29].

To relieve the requirement of expensive 3D labels, attempts are made to utilize more easily obtained annotations, *e.g.*, 2D landmarks and silhouettes [38, 42, 48, 51], ordinal depth relations [37], dense correspondences [7], or 3D skeletons [26]. To get rid of weak supervision, some take a step further by exploring temporal or multi-view images [23, 50] or prior knowledge such as poses with temporal consistency [55].

There has been some recent works using synthetic data for human body modeling, *e.g.*, dense correspondences estimation [65], depth estimation [49], 3D pose estimation [23, 35, 41, 47, 49], and 3D human reconstruction [63]. While these approaches show promising results, they need to render images under various synthetically designed conditions such as lighting and background. However, it is very challenging for such an approach to produce data (and hence the resulting trained model) that generalizes to real-world conditions. In contrast, [5, 43–46, 56] rely on various intermediate representations used for adjacent tasks such as keypoint, binary silhouettes, edges, and depth. Concretely, while insufficient data handicaps 3D human mesh estimation, tasks such as keypoint estimation have substantially more annotated data. This then leads to a situation where one can

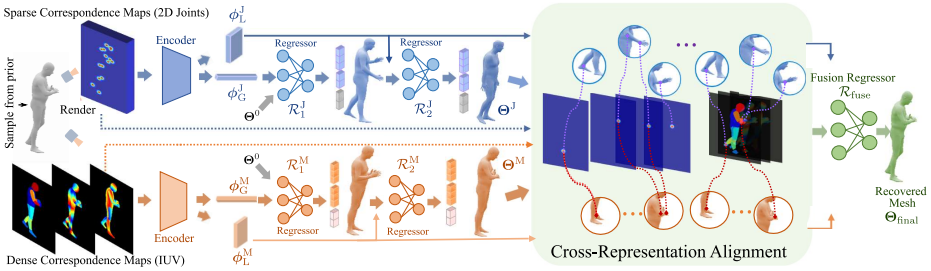

Fig. 2. Overview of the proposed pipeline with cross-representation alignment. For training, we generate paired data between 3D mesh and intermediate representations (*i.e.* 2D joints and IUV map).

expect intermediate representations for these tasks (*e.g.*, 2D keypoints estimation, binary silhouettes) to generalize better than the representation learned by standard mesh estimation models such as SPIN [19]. At test time on real data, all one needs to do is to compute these representations with off-the-shelf detectors and then infer with the trained intermediate-representation-to-mesh regressor.

Although the aforementioned synthesis-based methods regress the SMPL parameters directly from intermediate representations such as 2D keypoints, binary silhouettes, and depth, none of them successfully utilize synthetic dense correspondence maps (*i.e.*, IUV), which can provide richer and complementary information to 2D joints/edge/silhouette. While adding IUV to the representations may seem incremental, [43] acknowledge it is actually challenging due to the large domain gap between real IUV and synthetic IUV.

We propose cross-representation alignment (CRA) to address the large domain gap while employing dense intermediate representation in synthetic training to handle all the above considerations. Our critical insight is that all these representations may not be wholly consistent but come with complementary advantages. For instance, while 2D keypoints provide a robust sparse representation of the skeleton, dense correspondences (via UV maps) can help further finetune/finesse the final output (shown in Figure 1).

To this end, our proposed CRA fusion module comprises a trainable alignment scheme between the regressed mesh output and the evidential representations as part of an iterative feedback loop (shown in Figure 2). Unlike our counterparts [43, 46, 56] which simply concatenate the features from each representation and regress SMPL parameters iteratively. We instead exploit the complementary information among different representations by generating feedback based on alignment error between the mesh estimation and each representation. The alignment feedback is then forwarded into the following regressor inferring the final SMPL estimation. By introducing the feedback mechanism here, our proposed method can effectively exploit the complementary knowledge between both representations and adapt to their different characteristics, not only during training but also after deployment with real data.

To summarize, our key contributions are:

- We propose a novel synthetic-training pipeline successfully utilizing both sparse and dense representation by bridging the synthetic-to-real gap in dense correspondence via adaptive representation alignment.
- We capture complementary advantages in cross-modality with a trainable cross-representation fusion module that aligns the regressed mesh output with representation evidence as part of the iterative regression.
- We conduct extensive benchmarking on standard datasets and demonstrate competitive numbers with conventional evaluation metrics and protocols.

2 Related Work

Single-image human 3D pose/mesh estimation. The emergence of statistical body models such as SCAPE [1] and SMPL [31] makes it possible to represent the human body with low-dimensional parameters. Iterative optimization-based approaches have been leveraged to fit these parametric models to 2D observations such as keypoints [2, 36] and silhouettes [25]. These model-fitting approaches are time-consuming, sensitive to initialization, and difficult to tune. Recent advances are dominated by learning-based methods which regress a parametric model (*e.g.*, pose and shape parameters for SMPL [31]) or non-parametric model (*e.g.*, mesh vertices [20]) under the supervision of 3D labels. Several works learn 3D body mesh from image through intermediate representations, *e.g.*, surface keypoints [48, 51], silhouettes [38], body part segmentations [34], IUV maps [54,59,60], and 3D markers [58]. Others directly learn 3D body parameters from the input image [19]. Recent works have explored body kinematics [6, 53], pose augmentation, and pose probabilistic distributions [21] to boost performance. Self-attention and graph convolutional networks have also been used to learn relationships among vertices [27], body-parts [17, 66] to handle occlusions.
Weakly-supervised human 3D pose/mesh estimation. Several works take steps to leverage a variety of easily obtained clues, such as paired 2D landmarks and silhouettes [38, 42, 48, 51], ordinal depth relations [37], DensePose [7], 3D skeleton [26]. HMR [12] fits SMPL parameters to 2D ground-truth and utilizes adversarial learning to exploit unpaired 3D data to relieve the reliance on expensive 3D ground truth. Kundu *et al.* [22] learn human pose and shape with 2D evidence together with appearance consensus between pairs of images of the same person. Based on GHUM [52] as the parametric model, THUNDR [58] realizes weak-supervision via intermediate 3D marker representation.
Self-supervised human 3D pose/mesh estimation. Kundu *et al.* [22, 23] utilize temporal and multi-view images as pairs and background/foreground disentangling for self-supervision of human pose/mesh estimation. Multi-view self-supervised 3D pose estimation methods [18, 40, 50] usually require additional knowledge w.r.t. the scene and camera position or multi-view images. In the absence of multi-view video sequences and other views, geometric consistency [4], kinematics knowledge [24], and temporally consistent poses [55] have been explored for auxiliary prior self-supervision. HUND [57] utilizes in-the-wild images

and learns the mesh with differential rendering measures between predictions and image structures. Other synthesis-based methods generate 2D keypoints, silhouettes [43–45], and 3D skeleton [56] with existing MoCap data for training.

3 Method

3.1 Prerequisites

3D Human Mesh Parameterization: We parameterize the 3D human mesh using the Skinned Multi-Person Linear (SMPL) model. SMPL [31] is a parametric model providing independent body shape $\boldsymbol{\beta}$ and pose $\boldsymbol{\theta}$ representations with low-dimensional parameters (*i.e.*, $\boldsymbol{\beta} \in \mathbb{R}^{10}$ and $\boldsymbol{\theta} \in \mathbb{R}^{72}$). Pose parameters include global body rotation (3-DOF) and relative 3D rotations of 23 joints (23×3-DOF) in the axis-angle format. The shape parameters indicating individual heights and weights (among other parameters) are the first 10 coefficients of a PCA shape space. SMPL provides a differentiable kinematic function $\mathcal{S}$ from these pose/shape parameters to 6890 mesh vertices: $\boldsymbol{v} = \mathcal{S}(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \mathbb{R}^{6890\times 3}$. Besides, 3D joint locations for N_{J} joints of interest are obtained as $\boldsymbol{\jmath}^{\mathrm{3D}} = \mathcal{J}\boldsymbol{v}$, where $\mathcal{J} \in \mathbb{R}^{N_{\mathrm{J}}\times 6890}$ is a learned linear regression matrix.

Dense Human Body Representation: We use DensePose [8] to establish dense correspondence between the 2D image and the mesh surface behind clothes. It semantically defines 24 body parts as $\boldsymbol{I}$ to represent Head, Torso, Lower/Upper Arms, Lower/Upper Legs, Hands and Feet, where head, torso, and lower/upper limbs are partitioned into frontal-back parts to guarantee body parts are isomorphic to a plane. For UV parametrization, each body part index has a unique UV coordinate which is geometrically consistent. In this manner, with IUV representation each pixel can be projected back to vertices on the template mesh according to a predefined bijective mapping between the 3D surface space and the IUV space. We denote the IUV map as $[\boldsymbol{I}, \boldsymbol{U}, \boldsymbol{V}] \in \mathbb{R}^{3\times(P+1)\times H\times W}$, where $P = 24$ indicating 24 foreground body parts, H and W are the height and width of IUV map. The index channel is one-hot indicating whether it belongs to the background or specific body part: $\boldsymbol{I} \in \{0,1\}^{(P+1)\times H\times W}$. While $\boldsymbol{U}$ and $\boldsymbol{V}$ are independent channels containing the U, V values (ranging from 0 to 1) for corresponding body part [8]. IUV can be further reorganized as a more compact representation $\boldsymbol{M} = [\boldsymbol{M}^{\mathrm{I}}, \boldsymbol{M}^{\mathrm{U}}, \boldsymbol{M}^{\mathrm{V}}] \in \mathbb{R}^{3\times H\times W}$ which is convertible with the explicit one-hot IUV version mentioned above. With $h = 1, \ldots, H$ and $w = 1, \ldots, W$ as pixel position, we have $M^{I}_{hw} \in \{0, 1, \ldots, P\}$, where 0 indicates background and non-zero value indicates body part index. As at most one out of the $P + 1$ channels (background and body parts) has non-zero U/V values, the simplified $\boldsymbol{M}^{\mathrm{U}}$ and $\boldsymbol{M}^{\mathrm{V}}$ are represented by $M^{\mathrm{U}}_{hw} = U_{M^{\mathrm{I}}_{hw}hw}$, $M^{\mathrm{V}}_{hw} = V_{M^{\mathrm{I}}_{hw}hw}$.

3.2 Training Data Synthesis

We generate paired 2D representations and 3D meshes on-the-fly with SMPL. We utilize prior poses from the existing MoCap [3,41] datasets for diverse and realistic simulation. Body shape parameters are sampled from normal distribution

$\beta_n \sim \mathcal{N}(\mu_n, \sigma_n^2)(n = 1, \ldots, 10)$, where the mean and variance are empirically obtained from prior statistics [43] for generalization. We employ perspective projection with identity camera rotation $\boldsymbol{r} \in \mathbb{R}^{3\times3}$, dynamically sampled camera translation $\boldsymbol{t} \in \mathbb{R}^3$ as extrinsic parameters, and fixed focal length $\boldsymbol{f} \in \mathbb{R}^2$ as intrinsic parameters.

At each training step, the sampled $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are forwarded into SMPL model to obtain mesh vertex $\boldsymbol{v}$ and 3D joints $\boldsymbol{j}^{\mathrm{3D}}$. Then we project the 3D joints $\boldsymbol{j}^{\mathrm{3D}}$ to 2D joints $\boldsymbol{j}^{\mathrm{2D}}$, with sampled extrinsic and intrinsic camera parameters mentioned above: $\boldsymbol{j}^{\mathrm{2D}} = \boldsymbol{f}\Pi\,(\boldsymbol{r}\boldsymbol{j}^{\mathrm{3D}} + \boldsymbol{t})$, where Π denotes perspective projection. We normalize the $\boldsymbol{j}^{\mathrm{2D}}$ to be from -1 to 1, and denote normalized version as $\boldsymbol{j}^{\mathrm{2D}}$ in the following for simplification. With these camera parameters, we render the human mesh to 2D dense IUV based on an existing rendering method [39]. Specifically, we take predefined unique IUV value for each vertex on the SMPL model as a template, project the vertex IUV into 2D and then obtain a continuous 2D IUV map via rasterization and shading.

The 2D joints $\boldsymbol{j}^{\mathrm{2D}} \in \mathbb{R}^{N_{\mathrm{J}}\times2}$ are transformed into 2D Gaussian joint heatmaps $\boldsymbol{J} \in \mathbb{R}^{N_{\mathrm{J}}\times H\times W}$ as inputs to our neural networks. The IUV map with $\boldsymbol{M} \in \mathbb{R}^{3\times H\times W}$ is used as the other 2D representation. Note that we normalize the I channel in $\boldsymbol{M}$ to values between $[0, 1]$. For simplification, we subsequently denote the normalized version as $\boldsymbol{M}$. Finally, we have the synthesized paired data with 2D representations $\{\boldsymbol{j}^{\mathrm{2D}}, \boldsymbol{J}, \boldsymbol{M}\}$ and 3D mesh $\{\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{j}^{\mathrm{3D}}\}$.

3.3 Individual Coarse-to-fine Regression

Given the 2D representation (either $\boldsymbol{J}$ or $\boldsymbol{M}$), we first extract features with an encoder, then forward the features into the regressor, and predict the SMPL model with pose, shape, and camera parameters $\boldsymbol{\Theta} = \{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}\}$.

The encoder takes 2D representation as input and outputs features $\boldsymbol{\phi}_0 \in \mathbb{R}^{C_0\times H_0\times W_0}$. Before forwarding the features into the following regressor, we reduce the feature dimensions spatial-wisely and channel-wisely to maintain more global and local information. For global features, we use average-pooling to reduce spatial dimension and get $\boldsymbol{\phi}_{\mathrm{G}} = \mathrm{AvgPool}(\boldsymbol{\phi}_0) \in \mathbb{R}^{C_0\times1\times1}$. For fine-grained features, we use a multi-layer perceptron (MLP) for channel reduction and retain the spatial dimension the same:

$$\boldsymbol{\phi}_l = \begin{cases} \mathcal{P}_l(\boldsymbol{\phi}_{l-1}) & \text{if } l = 1 \\ \mathcal{P}_l(\boldsymbol{\phi}_{l-1} \oplus \boldsymbol{\phi}_0) & \text{if } l > 1, \end{cases} \tag{1}$$

where $\oplus$ denotes concatenation, $l = 1, \ldots, L$ is the perception layer, $\mathcal{P}_l$ indicates the l-th perceptron, and $\boldsymbol{\phi}_l \in \mathbb{R}^{C_l\times H_0\times W_0}$ with channel C_l monotonically decreasing. We denote the final output after MLP as $\boldsymbol{\phi}_{\mathrm{L}}$.

Taking the flattened feature $\boldsymbol{\phi}$ and initialized $\boldsymbol{\Theta}^0$ as input, the regressor $\mathcal{R}$ updates $\boldsymbol{\Theta} = \{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}\}$. Note here that we use continuous 6-dimensional representation [64] for optimization of joint rotation in $\hat{\boldsymbol{\theta}} \in \mathbb{R}^{24\times6}$ which can be converted to the discontinuous Euler rotation vectors. The predicted camera parameters for the standard weak-perspective projection are represented by

$\hat{\boldsymbol{\pi}} = [\hat{\pi}_{\mathrm{s}}, \hat{\boldsymbol{\pi}}_{\mathrm{t}}]$, where $\hat{\pi}_{\mathrm{s}} \in \mathbb{R}$ is the scale factor and $\hat{\boldsymbol{\pi}}_{\mathrm{t}} \in \mathbb{R}^2$ indicates translation. Similar to the standard iterative error feedback (IEF) procedure [12], we iteratively update the prediction $\boldsymbol{\Theta}$. For each representation ($\boldsymbol{J}$ and $\boldsymbol{M}$) stream, we have two regressors $\mathcal{R}_1$ and $\mathcal{R}_2$ estimating $\boldsymbol{\Theta}$ with global feature $\boldsymbol{\phi}_{\mathrm{G}}$ and fine-grained feature $\boldsymbol{\phi}_{\mathrm{L}}$ respectively:

$$\boldsymbol{\Theta}^{\mathrm{J}} = \mathcal{R}_2^{\mathrm{J}}(\boldsymbol{\phi}_{\mathrm{L}}^{\mathrm{J}}; \mathcal{R}_1^{\mathrm{J}}(\boldsymbol{\phi}_{\mathrm{G}}^{\mathrm{J}}; \boldsymbol{\Theta}^0)) \qquad \text{and} \qquad \boldsymbol{\Theta}^{\mathrm{M}} = \mathcal{R}_2^{\mathrm{M}}(\boldsymbol{\phi}_{\mathrm{L}}^{\mathrm{M}}; \mathcal{R}_1^{\mathrm{M}}(\boldsymbol{\phi}_{\mathrm{G}}^{\mathrm{M}}; \boldsymbol{\Theta}^0)), \tag{2}$$

where $\boldsymbol{\Theta}^{\mathrm{J}}$ and $\boldsymbol{\Theta}^{\mathrm{M}}$ are the parameter predictions for the 2D joints representation $\boldsymbol{J}$ and IUV representation $\boldsymbol{M}$ respectively.

3.4 Evidential Cross-Representation Alignment

To utilize the complementary information of both representations, we design a novel fusion module $\mathcal{R}_{\mathrm{fuse}}$ considering the misalignment between the prediction and the evidence from the intermediate representations (*i.e.*, 2D joints and IUV map). One observation is that the pose parameters are represented as relative rotations and kinematic trees where minor parameter differences can result in significant misalignment on 2D projections. Another observation is that the inferred 2D joints and IUV map are likely to be noisy and inconsistent in real scenarios. During testing, we can hardly distinguish which of the available 2D representations is more reliable, so we incorporate alignment between both pieces of evidence and both predictions.

Given $\boldsymbol{\Theta} = \{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}\}$ as prediction, SMPL takes $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$ to output 3D vertices $\hat{\boldsymbol{v}}$ and 3D joints $\hat{\boldsymbol{j}}^{3\mathrm{D}}$. Then with predicted camera parameters $\hat{\boldsymbol{\pi}}$, we have the reprojected 2D joints $\hat{\boldsymbol{j}}^{2\mathrm{D}} = \hat{\pi}_s \Pi\,(\hat{\boldsymbol{j}}^{3\mathrm{D}}) + \hat{\boldsymbol{\pi}}_t$ with orthographic projection function Π. We denote normalized version of $\hat{\boldsymbol{j}}^{2\mathrm{D}}$ as $\hat{\boldsymbol{j}}^{2\mathrm{D}}$ in the following for simplification. We also render the IUV map $\widehat{\boldsymbol{M}} \in \mathbb{R}^{3\times H_0 \times W_0}$ with $\hat{\boldsymbol{v}}$, $\hat{\boldsymbol{\pi}}$ and predefined unique IUV value for each vertex on the SMPL. Note that our projections and rendering techniques are differentiable.

To evaluate the misalignment on 2D joints, we have

$$\mathcal{D}_{\mathrm{J}}(\hat{\boldsymbol{j}}^{2\mathrm{D}}, \boldsymbol{j}^{2\mathrm{D}}) = \hat{\boldsymbol{j}}^{2\mathrm{D}} - \boldsymbol{j}^{2\mathrm{D}}, \tag{3}$$

where $\mathcal{D}_{\mathrm{J}}(\cdot,\cdot) \in \mathbb{R}^{N_{\mathrm{J}}\times 2}$ is a discrepancy vector which can also be seen as 2D joints pixel index offset between the prediction and the evidence. For misalignment between predicted IUV map $\widehat{\boldsymbol{M}} = [\widehat{\boldsymbol{M}}^{\mathrm{I}}, \widehat{\boldsymbol{M}}^{\mathrm{U}}, \widehat{\boldsymbol{M}}^{\mathrm{V}}]$ and evidential IUV map $\boldsymbol{M} \in \mathbb{R}^{3\times H\times W}$, we downsize $\boldsymbol{M}$ to be with $\mathbb{R}^{3\times H_0\times W_0}$. For simplicity, we use $\boldsymbol{M}$ to represent the downsized version from now on. The discrepancy map $\mathcal{D}_{\mathrm{M}}(\cdot,\cdot) \in \mathbb{R}^{H_0\times W_0}$ can be obtained:

$$\begin{aligned}\mathcal{D}_{\mathrm{M}}(\widehat{\boldsymbol{M}}, \boldsymbol{M}) = \frac{|\widehat{\boldsymbol{M}}^{\mathrm{I}} - \boldsymbol{M}^{\mathrm{I}}|}{|\widehat{\boldsymbol{M}}^{\mathrm{I}} - \boldsymbol{M}^{\mathrm{I}}|_{\mathrm{d}} + \epsilon} &+ \sum_{p=1}^{P} [\mathbf{1}(\widehat{\boldsymbol{M}}^{\mathrm{I}} = \frac{p}{P}) \odot \widehat{\boldsymbol{M}}^{\mathrm{U}} - \mathbf{1}(\boldsymbol{M}^{\mathrm{I}} = \frac{p}{P}) \odot \boldsymbol{M}^{\mathrm{U}}] \\ &+ \sum_{p=1}^{P} [\mathbf{1}(\widehat{\boldsymbol{M}}^{\mathrm{I}} = \frac{p}{P}) \odot \widehat{\boldsymbol{M}}^{\mathrm{V}} - \mathbf{1}(\boldsymbol{M}^{\mathrm{I}} = \frac{p}{P}) \odot \boldsymbol{M}^{\mathrm{V}}],\end{aligned} \tag{4}$$

where the $|\cdot|$ indicates ℓ_1 norm, $(\cdot)_{\mathrm{d}}$ indicates detachment from gradients, $\epsilon = 1e^{-5}$ is to prevent the denominator to be zero; thus the first term corresponds to a differentiable version of the indicator function $\mathbf{1}(\widehat{\boldsymbol{M}}^{\mathrm{I}}{=}\widetilde{\boldsymbol{M}}^{\mathrm{I}})$. In the second and third terms, $P = 24$ indicates the 24 body parts, $\odot$ denotes element-wise multiplication, and the indicator function $\mathbf{1}$ judges whether $\widetilde{\boldsymbol{M}}^{\mathrm{I}}$ or $\widehat{\boldsymbol{M}}^{\mathrm{I}}$ corresponds to specific body part p, which is normalized here as $\frac{p}{P}$.

To simplify the notations, from this point on, we refer to $\hat{\boldsymbol{j}}^{\mathrm{2D}}$ as $\hat{\boldsymbol{j}}$, we have $\{\hat{\boldsymbol{j}}^{\mathrm{J}}, \widehat{\boldsymbol{M}}^{\mathrm{J}}\}$ and $\{\hat{\boldsymbol{j}}^{\mathrm{M}}, \widehat{\boldsymbol{M}}^{\mathrm{M}}\}$ corresponding to $\boldsymbol{\Theta}^{\mathrm{J}}$ and $\boldsymbol{\Theta}^{\mathrm{M}}$ respectively. Then we have $\boldsymbol{D}_{\mathrm{J}}^{\mathrm{J}} = \mathcal{D}_{\mathrm{J}}(\hat{\boldsymbol{j}}^{\mathrm{J}}, \boldsymbol{j}^{\mathrm{2D}})$ and $\boldsymbol{D}_{\mathrm{J}}^{\mathrm{M}} = \mathcal{D}_{\mathrm{J}}(\hat{\boldsymbol{j}}^{\mathrm{M}}, \boldsymbol{j}^{\mathrm{2D}})$ as the 2D joints misalignment between the two predictions and the evidence. And $\boldsymbol{D}_{\mathrm{M}}^{\mathrm{J}} = \mathcal{D}_{\mathrm{M}}(\widehat{\boldsymbol{M}}^{\mathrm{J}}, \boldsymbol{M})$ and $\boldsymbol{D}_{\mathrm{M}}^{\mathrm{M}} = \mathcal{D}_{\mathrm{M}}(\widehat{\boldsymbol{M}}^{\mathrm{M}}, \boldsymbol{M})$ as the IUV misalignment between the two predictions and the evidences. All these misalignment representations are flattened and then taken as input of $\mathcal{R}_{\mathrm{fuse}}$ along with the flattened features $\boldsymbol{\phi}_{\mathrm{L}}^{\mathrm{J}}$ and $\boldsymbol{\phi}_{\mathrm{L}}^{\mathrm{M}}$:

$$\boldsymbol{\Theta}^{\mathrm{final}} = \mathcal{R}_{\mathrm{fuse}}(\boldsymbol{D}_{\mathrm{J}}^{\mathrm{J}}, \boldsymbol{D}_{\mathrm{M}}^{\mathrm{J}}, \boldsymbol{\phi}_{\mathrm{L}}^{\mathrm{J}}, \boldsymbol{D}_{\mathrm{J}}^{\mathrm{M}}, \boldsymbol{D}_{\mathrm{M}}^{\mathrm{M}}, \boldsymbol{\phi}_{\mathrm{L}}^{\mathrm{M}}; \boldsymbol{\Theta}^{\mathrm{J}}, \boldsymbol{\Theta}^{\mathrm{M}}), \tag{5}$$

where $\boldsymbol{\Theta}^{\mathrm{final}}$ is the final prediction initialized with both $\boldsymbol{\Theta}^{\mathrm{J}}$ and $\boldsymbol{\Theta}^{\mathrm{M}}$. Note that each step of the fusion module is differentiable, *i.e.*, maintaining the gradients so that the following loss function is able to penalize misalignment and correct the precedent prediction from $\mathcal{R}_1^{\mathrm{J}}$, $\mathcal{R}_2^{\mathrm{J}}$, $\mathcal{R}_1^{\mathrm{M}}$, $\mathcal{R}_2^{\mathrm{M}}$ during training.

3.5 Loss Function

As described in Section 3.4, from $\boldsymbol{\Theta}^{\mathrm{final}}$ we can obtain predicted vertices $\hat{\boldsymbol{v}}$, 3D joints $\hat{\boldsymbol{j}}^{\mathrm{3D}}$, and project to 2D joints $\hat{\boldsymbol{j}}^{\mathrm{2D}}$. We have prediction and supervision in terms of vertices, 2D joints, 3D joints and SMPL parameters respectively. To balance among these parts, we make the loss weights learnable using homoscedastic uncertainty as in prior works [14, 43]:

$$\begin{aligned}
&\mathcal{L}_{\mathrm{reg}}(\hat{\boldsymbol{v}}, \hat{\boldsymbol{j}}^{\mathrm{2D}}, \hat{\boldsymbol{j}}^{\mathrm{3D}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}, \boldsymbol{v}, \boldsymbol{j}^{\mathrm{2D}}, \boldsymbol{j}^{\mathrm{3D}}, \boldsymbol{\theta}, \boldsymbol{\beta}) \\
=&\frac{\mathcal{L}_2(\hat{\boldsymbol{v}}, \boldsymbol{v})}{\sigma_{\mathrm{v}}^2} + \frac{\mathcal{L}_2(\hat{\boldsymbol{j}}^{\mathrm{2D}}, \boldsymbol{j}^{\mathrm{2D}})}{\sigma_{\mathrm{j2D}}^2} + \frac{\mathcal{L}_2(\hat{\boldsymbol{j}}^{\mathrm{3D}}, \boldsymbol{j}^{\mathrm{3D}})}{\sigma_{\mathrm{j3D}}^2} \\
&+ \frac{\mathcal{L}_2([\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}], [\boldsymbol{\theta}, \boldsymbol{\beta}])}{\sigma_{\mathrm{SMPL}}^2} + \log(\sigma_{\mathrm{v}}\sigma_{\mathrm{j2D}}\sigma_{\mathrm{j3D}}\sigma_{\mathrm{SMPL}}),
\end{aligned} \tag{6}$$

where $\mathcal{L}_2$ denotes the mean square error (MSE), and σ_{v}, σ_{j2D}, σ_{j3D} and σ_{SMPL} indicates weights for vertex, 2D joints, 3D joints, SMPL parameters which are adaptively adjusted during training.

Auxiliary Refinement. Our framework can naturally refine the network with available in-the-wild images. Given an image, we use an existing off-the-shelf detector to obtain IUV map $\boldsymbol{M}$ and 2D joints $\boldsymbol{j}^{\mathrm{2D}}$. The IUV map is downsampled and the 2D joints are processed to Gaussian heatmaps $\boldsymbol{J}$. We take $\{\boldsymbol{M}, \boldsymbol{J}\}$ as input, forward through our network, and output the final prediction $\boldsymbol{\Theta}$. As described in Section 3.4, we obtain the reprojected $\hat{\boldsymbol{j}}^{\mathrm{2D}}$ and rendered $\widehat{\boldsymbol{M}}$ in

a differentiable manner. Given $\mathcal{D}_{\mathrm{M}}$ defined in Equation 4, the refinement loss function is thus computed as:

$$\mathcal{L}_{\text{refine}}(\hat{\boldsymbol{j}}^{\text{2D}}, \widehat{\boldsymbol{M}}, \boldsymbol{j}^{\text{2D}}, \boldsymbol{M}) = \mathcal{L}_2(\hat{\boldsymbol{j}}^{\text{2D}}, \boldsymbol{j}^{\text{2D}}) + \mathcal{D}_{\mathrm{M}}(\widehat{\boldsymbol{M}}, \boldsymbol{M}), \tag{7}$$

4 Experiments

4.1 Datasets

Training data. To generate synthetic training data, we sample SMPL pose parameters from the training sets of UP-3D [25], 3DPW [32], and the five training subjects of Human3.6M [11] (S1, S5, S6, S7, S8). The sampling of shape parameters follows the procedure of prior work [43].
Evaluation data. We report evaluation results on both indoor and outdoor datasets, including 3DPW [32], MPI-INF-3DHP [33], and Human3.6M [11] (Protocols 1 and 2 [12] with subjects S9, S11). For 3DPW, we report the mean per joint position error (MPJPE), mean per joint position error after rigid alignment with Procrustes analysis (PMPJPE), and after-scale correction [43] for pose estimation, and per-vertex error (PVE) for shape estimation. For MPI-INF-3DHP, we report metrics after rigid alignment, including PMPJPE, percentage of correct keypoints (PCK) thresholded at 150mm, and the area under the curve (AUC) over a range of PCK thresholds [33]. For Human3.6M, we report MPJPE and PMPJPE on protocols 1 and 2 using the H3.6M joints definition.

4.2 Implementation Details

Synthetic data preprocessing and augmentation: We generate paired data on the fly with details described in Section 3.2. We follow the hyperparameters in [43] for SMPL shape and camera translation sampling. We use $N_{\mathrm{J}} = 17$ COCO joints to extract 3D joints from the SMPL model and then project to 2D joints representation. The vertices $\boldsymbol{v}$ are randomly perturbed within $[-10\text{mm}, 10\text{mm}]$ for augmentation. From perturbed vertices and sampled camera parameters, we render 2D IUV map $\boldsymbol{M}$ based on Pytorch3D [39]. We detect the foreground body area on 2D IUV and crop around the foreground area with a scale of 1.2 around the bounding box, which is unified for consistency between training and testing. We crop both IUV $\boldsymbol{M}$ and joints heatmaps $\boldsymbol{J}$ and then resize to the target size with $H = 256$, $W = 256$. To simulate noise and discrepancy between 2D joints and IUV prediction, we do a series of probabilistic augmentations, including randomly masking one of the six body parts (same as PartDrop in [60]), randomly masking one of the six body parts (head, torso, left/right arm, left/right leg) on IUV map, randomly occluding the IUV map with a dynamically-sized rectangle, and randomly perturbing the 2D joints position.
Architecture: We use ResNet-18 [10] as encoder and the size of the output $\boldsymbol{\phi}_0$ is $C_0 = 512$, $H_0 = 8$, $W_0 = 8$. Through average pooling we get $\boldsymbol{\phi}_{\mathrm{G}}$ with size $512 \times 1 \times 1$. Each perceptron $\mathcal{P}_l$ in the MLP consists of Conv1D and ReLU operations with $L = 3$ layers in total. The MLP reduce the feature channels to

Method	2D Superv.	Auxiliary requirements: image pairs	multi-view imagery	temporal prior	Protocol # 1 MPJPE↓	Protocol # 1 PMPJPE↓	Protocol # 2 MPJPE↓	Protocol # 2 PMPJPE↓
*HMR (unpaired) [12]	✓	✗	✗	✗	106.84	67.45		66.5
*SPIN (unpaired) [19]	✓	✗	✗	✗	-	-	-	62.0
*Kundu *et al.* [22]	✓	✓	✗	✗	**86.4**			**58.2**
*THUNDER [58]	✓	✗	✗	✗	87.0	62.2	83.4	59.7
Kundu *et al.* [24]	✗	✓	✗	✗	-	-	-	89.4
Kundu *et al.* [23]	✗	✓	✓	✗	-	-	-	85.8
*Kundu *et al.* [22]	✗	✓	✓	✗	102.1	-	-	74.1
CanonPose [50]	✗	✗	✓	✗	**81.9**	-	-	53
Yu *et al.* [55]	✗	✗	✗	✓	-	-	92.4	**52.3**
*Song *et al.* [46]	✗	✗	✗	✗	-	-	-	56.4
*STRAP [43]	✗	✗	✗	✗	87.0	59.3	83.1	55.4
*HUND [57]	✗	✗	✗	✗	91.8	66.0	-	-
*Skeleton2Mesh [56]	✗	✗	✗	✗	87.1	**55.4**	-	-
*Ours (*synthesis only*)	✗	✗	✗	✗	87.1	58.2	81.3	54.8
*Ours (*w/ refinement*)	✗	✗	✗	✗	**84.3**	57.8	**81.0**	**53.9**

Table 1. Comparison of our method with weakly supervised and self-supervised SOTA in terms of MPJPE and PMPJPE (both in mm) on the H3.6M Protocol #1 and Protocol #2 test sets. * indicates methods that can estimate more than 3D pose.

$[C_1, C_2, C_3] = [256, 64, 8]$ progressively, and produces the feature vector $\boldsymbol{\phi}_{\mathrm{L}}$ with size $8 \times 8 \times 8$. Each regression network for $\{\mathcal{R}_1^{\mathrm{J}}, \mathcal{R}_2^{\mathrm{J}}, \mathcal{R}_1^{\mathrm{M}}, \mathcal{R}_2^{\mathrm{M}}\}$ consists of two fully connected layers with 512 neurons each, followed by an output layer with 157 neurons ($\boldsymbol{\Theta} = \{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}\} \in \mathbb{R}^{24\times6+10+3}$ as explained in Section 3.3). Taking the input vector with dimension $2 \times (C_L \times H_0 \times W_0 + 3 \times H_0 \times W_0 + 2 \times N_{\mathrm{J}}) = 1540$, the regression network for $\mathcal{R}_{\mathrm{fuse}}$ consists of two fully connected layers with 1540 neurons each, followed by an output layer with 157 neurons.

Training: With the final prediction $\boldsymbol{\Theta}_{\mathrm{final}}$, we use Equation 6 as a loss function to train the whole network in an end-to-end fashion. We use Adam [15] optimizer to train for 30 epochs with a learning rate of $1e^{-4}$ and a batch size of 128. On the image, we predict 2D joints and IUV maps using the off-the-shelf Keypoint-RCNN [9] and DensePose [8] models. For auxiliary refinement, we use RGB images from the corresponding training set when testing on the Human3.6M, 3DPW, and MPI-INF-3DHP. We use Adam to train for ten epochs with a learning rate of $1e^{-6}$ and a batch size of 128 for auxiliary refinement.

Testing: We infer 2D joints on the testing images with the pretrained Keypoint-RCNN [9] with ResNet-50 backbone. We obtain the IUV prediction with pretrained DensePose-RCNN [9] with ResNet-101 backbone. Since 3DPW test images may have multiple persons, we use the same protocol as [19] to get the bounding box for the target person by using the scale and center information and get the 2D representations with maximum IOU with the target bounding box. We crop both the IUV maps and 2D joints heatmaps with a scale of 1.2 before forwarding them to the network for 3D mesh inference.

	Method	PVE↓	MPJPE↓	MPJPE-SC↓	PMPJPE↓
Full Superv.	HMR [12]	139.3	116.5	-	72.6
	VIBE [16]	113.4	113.4	-	56.5
	PyMAF [61]	110.1	92.8	-	58.9
Weak Superv.	HMR (unpaired) [12]	-	-	126.3	92.0
	Kundu *et al.* [22]	-	153.4	-	89.8
	THUNDER [58]	-	87.8	-	59.9
Self Superv.	Kundu *et al.* [22]	-	187.1	-	102.7
	STRAP [43]	131.4	118.3	99.0	66.8
	HUND [57]	-	<u>90.4</u>	-	63.5
	STRAP V2 [45]	-	-	90.9	61.0
	STRAP V3 [44]	-	-	84.7	59.2
	Song *et al.* [46]	-	-	-	**55.9**
	Ours (*synthesis only*)	117.4	91.1	<u>80.8</u>	<u>56.3</u>
	Ours (*w/ refinement*)	115.3	**89.1**	**79.0**	**55.9**

Table 2. A comparison with fully/weakly/self-supervised SOTA methods in terms of PVE, MPJPE, MPJPE-SC, and PMPJPE (all in mm) on the 3DPW test dataset.

4.3 Quantitative Results

Human3.6M: We evaluate our method on the Human3.6M [11] test dataset (both Protocol #1 and Protocol #2) and compare our method with SOTA weakly supervised methods and self-supervised methods in Table 1. Note that the weakly supervised methods utilized paired images and 2D ground-truth such as 2D joints for supervision during training. And some self-supervised methods use auxiliary clues such as image pairs in video sequences, multi-view images, or prior knowledge of human keypoint positions on temporal sequences. Without reliance on either of these prerequisites, our method shows very competitive results compared with the prior arts with auxiliary refinement. Among the methods not requiring auxiliary clues, *e.g.* temporal or multi-view imagery, we achieve the best results in 3D pose estimation metrics-MPJPE Protocol #1, MPJPE, and PMPJPE on Protocol #2 of the Human3.6M test set.

3DPW: On the test set of 3DPW [32], we calculate PVE as shape evaluation metric and MPJPE, PMPJPE, MPJPE-SC [43] as pose evaluation metrics. From the comparisons in Table 2, we note that our method outperforms the prior arts, including those trained with 3D ground-truth (*i.e.*, full supervision) and 2D ground-truth (*i.e.*, weak supervision), on all metrics for pose evaluation. Although we do not rely on any annotated data, our method achieves results on shape estimation comparable to the prior arts trained with 3D annotation.

MPI-INF-3DHP: On the test set of MPI-INF-3DHP [33], we consider the usual metrics PCK, AUC, and PMPJPE after rigid alignment, to evaluate the 3D pose estimation. As shown in Table 3, other methods heavily rely on the related human image dataset for training, and some have additional requirements on multi-view images (*i.e.*, Human3.6M) and continuous images in temporal sequence (*i.e.*, YouTube videos). In contrast, our method has no such requirements and yet achieves better results on PCK than the prior arts (including weakly supervised methods). With access to the images, we can refine the network with a 0.9 mm improvement in PMPJPE. Compared with the methods relying on

Method	Images Used	PCK↑	AUC↑	PMPJPE↓
*HMR (unpaired) [12]	H36M+3DHP	77.1	40.7	113.2
Kundu *et al.* [24]	H36M+3DHP	79.2	43.4	99.2
Kundu *et al.* [23]	H36M+YTube	**83.2**	**58.7**	97.6
CanonPose [50]	H36M+YTube	77.0	-	**70.3**
Yu *et al.* [55]	3DHP	86.2	51.7	-
*Skeleton2Mesh [56]	3DHP	87.0	50.8	87.4
*SPIN (unpaired) [19]	3DHP	87.0	48.5	80.4
*Ours (*synthesis only*)	**None**	89.4	54.0	80.2
*Ours (*w/ refinement*)	3DHP	**89.7**	**55.0**	**79.1**

Table 3. Comparison with SOTA methods in terms of PCK, AUC, and PMPJPE (mm) after rigid alignment on the MPI-INF-3DHP test dataset. * indicates methods that can estimate more than 3D pose. Methods in the top half require training images paired with 2D ground-truth. Methods in the bottom half do not.

Representation	Regressor	Fusion	PVE↓	PMPJPE↓
[1]J2D	$\mathcal{R}_1$ $\mathcal{R}_1$ $\mathcal{R}_1$	-	181.3	75.2
[2]IUV	$\mathcal{R}_1$ $\mathcal{R}_1$ $\mathcal{R}_1$	-	167.2	83.1
[3]J2D & IUV	$\mathcal{R}_1$ $\mathcal{R}_1$ $\mathcal{R}_1$	input $\oplus$	121.3	60.1
[4]J2D & IUV	$\{\mathcal{R}_1\ \mathcal{R}_1\}^{\times 2}$	$\mathcal{R}_{\text{fuse}}$	120.8	61.0
[5]J2D & IUV	$\{\mathcal{R}_1\ \mathcal{R}_2\}^{\times 2}$	$\mathcal{R}_{\text{fuse}}$	117.7	59.6
[6]J2D & IUV	$\{\mathcal{R}_1\ \mathcal{R}_2\}^{\times 2}$	${}^{\triangleleft}\mathcal{R}_{\text{fuse}}$	118.6	58.2
[7]J2D & IUV	$\{\mathcal{R}_1\ \mathcal{R}_2\}^{\times 2}$	${}^{\triangleright\triangleleft}\mathcal{R}_{\text{fuse}}$	**117.4**	**56.3**

Table 4. Ablations of one/two representations, concatenation fusion, two-stream fusion with regressor $\mathcal{R}_3$, and the evidential representation alignment on the 3DPW test dataset in terms of PVE and PMPJPE (mm). Here ${}^{\triangleleft}\mathcal{R}$, ${}^{\triangleright\triangleleft}\mathcal{R}$ denotes the regressor taking misalignment of its preceding regressor prediction in terms of ${}^{\triangleleft}$the other/${}^{\triangleright\triangleleft}$both representation(s) as additional input. Note: no refinement applied for comparison.

both temporal and multi-view images [23, 50], our method achieves state-of-the-art PCK and very competitive AUC and PMPJPE without any requirements of images. Notably, we do not use any prior information of MPI-INF-3DHP during synthetic training but still achieve very competitive performance on MPI-INF-3DHP with model only trained with synthetic data. This demonstrates the superiority of our method's generalization ability to unseen in-the-wild data.

Ablations: In Table 4, we study the efficacy of our cross-representation alignment, where $\oplus$ denotes concatenate two representations as input of the encoder for fusion. From line 1 to line 3, we note that using the complementary information of 2D joints and IUV is better than using only one. The bottom half shows the results under our two-stream fusion pipeline, demonstrating the efficacy of our alignment module. The comparison between line 4 and line 5 shows that separate regressors taking features with different scales achieve better results than iterative regression with $\mathcal{R}_1$ only taking features with size $C_0 \times 1 \times 1$. And the incorporation of our evidential representation alignment scheme (discrepancy vector/map (Equation 3/4) between the preceding regressor's prediction and the evidence as an additional input of the regressor) achieves further improvement

	Body part occlusion prob.	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
PVE↓	IUV	163.4	166.1	169.0	171.7	174.6	177.0	180.1	182.7	185.5
	IUV + J2D (wo/ CRA)	138.8	143.6	145.8	148.3	150.7	153.0	155.7	157.9	160.3
	IUV + J2D (w/ CRA)	**118.1**	**118.7**	**119.4**	**120.1**	**120.7**	**121.4**	**122.0**	**122.7**	**123.4**
PMPJPE↓	IUV	92.9	94.9	97.0	99.0	101.1	102.9	105.1	106.9	109.1
	IUV + J2D (wo/ CRA)	61.8	62.6	64.3	66.2	68.0	69.8	71.7	73.3	75.1
	IUV + J2D (w/ CRA)	**56.8**	**57.3**	**57.8**	**58.2**	**58.7**	**59.2**	**59.6**	**60.1**	**60.6**
	Remove 2D joints prob.	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
PVE↓	J2D	185.0	193.9	204.2	237.8	273.7	306.8	339.6	370.9	402.1
	J2D + IUV (wo/ CRA)	150.5	164.8	181.9	201.9	222.6	246.6	270.1	294.0	318.7
	J2D + IUV (w/ CRA)	**127.4**	**139.8**	**153.9**	**170.2**	**188.9**	**210.1**	**232.9**	**258.0**	**284.2**
PMPJPE↓	J2D	88.5	98.8	122.5	145.8	169.4	189.7	208.5	224.6	239.0
	J2D + IUV (wo/ CRA)	68.0	78.4	90.0	102.8	114.9	127.5	138.7	148.3	156.3
	J2D + IUV (w/ CRA)	**63.9**	**72.7**	**82.5**	**92.8**	**104.1**	**115.7**	**128.1**	**140.5**	**153.6**

Table 5. Comparisons of PVE and PMPJPE (both in mm) when adding noise on IUV/2D joints representations of 3DPW test images. We study the performances when using one/two representations and using two representations with and without CRA.

(line 6 over line 7). We can see that utilizing discrepancy on both representations before $\mathcal{R}_{\text{fuse}}$ achieves the best result.

To study the efficiency of our proposed cross-representation alignment, we further simulate the extremely challenging conditions by adding noise on the inferred 2D joints and IUV representations. On the IUV map, we simulate the occlusion cases by masking out one of the six coarse body parts (head, torso, left/right arm, left/right leg) with increasing probability. For 2D joints, we remove the key joints(*i.e.*, left and right elbow, wrist, knee, ankle) with increasing probability. From the comparisons in Table 5, we can see that the combination of 2D joints and IUV can outperform IUV only on both shape and pose evaluations. Notably, our proposed cross-representation alignment (w/ CRA) outperforms the baseline (wo/ CRA) by a large margin, especially for the cases with severe noise.

4.4 Qualitative Results

Qualitative examples are given in Figure 3(a). We compare our proposed CRA (row 4) with typical concatentation taking 2D joints and IUV as input (row 2), and the baseline of CRA with no alignment applied (row 3). From the highlighted part we can see that our method with alignment module achieves much better shape estimation as well as pose estimation especially on joints such as wrist and knees. Notably the visualization is on images selected from SSP-3D [43] and MPI-INF-3DHP test set of which we do not utilize any prior knowledge. The results demonstrate the robustness and generalization ability of our proposed method to unseen in-the-wild data. We observe that for a small number of cases it could be difficult for CRA to recover from errors existing in all input immediate representations (*e.g.*, no detection on the lower body in both sparse and dense correspondences).

Auxiliary reconstruction: Our two-stream pipeline enables utilization of the encoded features ϕ of one representation to reconstruct another represen-

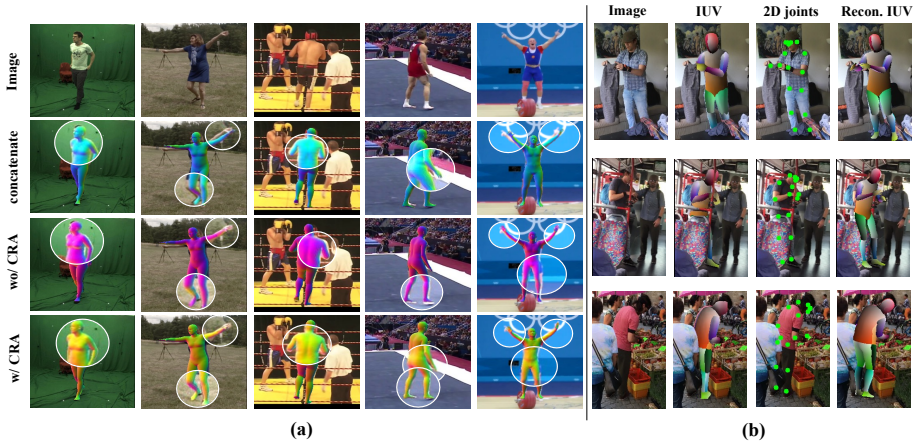

Fig. 3. (a) Comparison of qualitative results on human mesh estimation: taking 2D joints and IUV as input and processing with concatenation, two-stream fusion with and without GRU. (b) Visualization of IUV, 2D joints and reconstructed IUV from 2D joints on 3DPW test set. Note IUV (col 1) is visualized in HSV color space, which is predicted from pretrained Densepose-RCNN. 2D joints (col 2) are predicted from Keypoint-RCNN. IUV reconstructed from the 2D joints by the decoder (col 3) is trained together with CRA.

tation (*e.g.* from 2D joints to IUV map) at the same time while recovering the human mesh. We use a symmetric version of encoder as the decoder for each representation and employ the same loss function as [61] for IUV reconstruction with the synthetic IUV as supervision during training. From Figure 3(b), we note that the IUV prediction from off-the-shelf detector (trained with annotation) is occasionally sensitive to occlusion. While our recovered IUV trained with synthetic data can generalize to occlusion and more robust to ambiguous area in RGB image.

5 Conclusion

We propose a novel human mesh recovery framework relying only on synthetically generated intermediate representations based on pose priors. We design a Cross-Representation Alignment module to exploit complementary features from these intermediate modalities by enforcing consistency between predicted mesh parameters and input representations. Experimental results on popular benchmark datasets demonstrate the efficacy and generalizability of this framework.

References

1. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: shape completion and animation of people. In: ACM SIGGRAPH 2005 Papers, pp. 408–416 (2005) 4
2. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: European conference on computer vision. pp. 561–578. Springer (2016) 1, 4
3. C: Mocap. In: mocap. cs. cmu (2003) 5
4. Chen, C.H., Tyagi, A., Agrawal, A., Drover, D., Stojanov, S., Rehg, J.M.: Unsupervised 3d pose estimation with geometric self-supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5714–5724 (2019) 4
5. Clever, H.M., Grady, P., Turk, G., Kemp, C.C.: Bodypressure-inferring body pose and contact pressure from a depth image. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022) 2
6. Georgakis, G., Li, R., Karanam, S., Chen, T., Košecká, J., Wu, Z.: Hierarchical kinematic human mesh recovery. In: European Conference on Computer Vision. pp. 768–784. Springer (2020) 4
7. Guler, R.A., Kokkinos, I.: Holopose: Holistic 3d human reconstruction in-the-wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10884–10894 (2019) 2, 4
8. Güler, R.A., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7297–7306 (2018) 5, 10
9. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017) 10
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016) 9
11. 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) 9, 11
12. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). pp. 7122–7131 (2018) 1, 4, 7, 9, 10, 11, 12
13. Karanam, S., Li, R., Yang, F., Hu, W., Chen, T., Wu, Z.: Towards contactless patient positioning. IEEE transactions on medical imaging **39**(8), 2701–2710 (2020) 1
14. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7482–7491 (2018) 8
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 10
16. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5253–5263 (2020) 11
17. Kocabas, M., Huang, C.H.P., Hilliges, O., Black, M.J.: Pare: Part attention regressor for 3d human body estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11127–11137 (October 2021) 4

18. Kocabas, M., Karagoz, S., Akbas, E.: Self-supervised learning of 3d human pose using multi-view geometry. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1077–1086 (2019) 4
19. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2252–2261 (2019) 1, 3, 4, 10, 12
20. Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4501–4510 (2019) 4
21. Kolotouros, N., Pavlakos, G., Jayaraman, D., Daniilidis, K.: Probabilistic modeling for human mesh recovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11605–11614 (October 2021) 4
22. Kundu, J.N., Rakesh, M., Jampani, V., Venkatesh, R.M., Babu, R.V.: Appearance consensus driven self-supervised human mesh recovery. In: European Conference on Computer Vision. pp. 794–812. Springer (2020) 4, 10, 11
23. Kundu, J.N., Seth, S., Jampani, V., Rakesh, M., Babu, R.V., Chakraborty, A.: Self-supervised 3d human pose estimation via part guided novel image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6152–6162 (2020) 2, 4, 10, 12
24. Kundu, J.N., Seth, S., Rahul, M., Rakesh, M., Radhakrishnan, V.B., Chakraborty, A.: Kinematic-structure-preserved representation for unsupervised 3d human pose estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11312–11319 (2020) 4, 10, 12
25. 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) 4, 9
26. Li, J., Xu, C., Chen, Z., Bian, S., Yang, L., Lu, C.: Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3383–3393 (June 2021) 2, 4
27. Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1954–1963 (2021) 4
28. Liu, S., Song, L., Xu, Y., Yuan, J.: Nech: Neural clothed human model. In: 2021 International Conference on Visual Communications and Image Processing (VCIP). pp. 1–5. IEEE (2021) 1
29. Liu, S., Huang, X., Fu, N., Li, C., Su, Z., Ostadabbas, S.: Simultaneously-collected multimodal lying pose dataset: Enabling in-bed human pose monitoring. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022) 2
30. Loper, M., Mahmood, N., Black, M.J.: Mosh: Motion and shape capture from sparse markers. ACM Transactions on Graphics (TOG) **33**(6), 1–13 (2014) 2
31. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM transactions on graphics (TOG) **34**(6), 1–16 (2015) 1, 4, 5
32. 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) 9, 11

33. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 2017 international conference on 3D vision (3DV). pp. 506–516. IEEE (2017) 9, 11
34. Omran, M., Lassner, C., Pons-Moll, G., Gehler, P., Schiele, B.: Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In: 2018 international conference on 3D vision (3DV). pp. 484–494. IEEE (2018) 4
35. Patel, P., Huang, C.H.P., Tesch, J., Hoffmann, D.T., Tripathi, S., Black, M.J.: Agora: Avatars in geography optimized for regression analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13468–13478 (2021) 2
36. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10975–10985 (2019) 4
37. Pavlakos, G., Zhou, X., Daniilidis, K.: Ordinal depth supervision for 3d human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7307–7316 (2018) 2, 4
38. Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K.: Learning to estimate 3d human pose and shape from a single color image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 459–468 (2018) 2, 4
39. 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) 6, 9
40. Rhodin, H., Salzmann, M., Fua, P.: Unsupervised geometry-aware representation for 3d human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 750–767 (2018) 4
41. Rogez, G., Schmid, C.: Mocap-guided data augmentation for 3d pose estimation in the wild. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 3108–3116 (2016) 2, 5
42. Rong, Y., Liu, Z., Li, C., Cao, K., Loy, C.C.: Delving deep into hybrid annotations for 3d human recovery in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5340–5348 (2019) 2, 4
43. Sengupta, A., Budvytis, I., Cipolla, R.: Synthetic training for accurate 3d human pose and shape estimation in the wild. In: BMVC (2020) 2, 3, 5, 6, 8, 9, 10, 11, 13
44. Sengupta, A., Budvytis, I., Cipolla, R.: Hierarchical kinematic probability distributions for 3d human shape and pose estimation from images in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11219–11229 (October 2021) 2, 5, 11
45. Sengupta, A., Budvytis, I., Cipolla, R.: Probabilistic 3d human shape and pose estimation from multiple unconstrained images in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16094–16104 (June 2021) 2, 5, 11
46. Song, J., Chen, X., Hilliges, O.: Human body model fitting by learned gradient descent. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16. pp. 744–760. Springer (2020) 2, 3, 10, 11
47. Song, L., Yu, G., Yuan, J., Liu, Z.: Human pose estimation and its application to action recognition: A survey. Journal of Visual Communication and Image Representation **76**, 103055 (2021) 2

48. Tan, J., Budvytis, I., Cipolla, R.: Indirect deep structured learning for 3d human body shape and pose prediction. In: British Machine Vision Conference 2017, BMVC 2017 (2017) 2, 4
49. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 109–117 (2017) 2
50. Wandt, B., Rudolph, M., Zell, P., Rhodin, H., Rosenhahn, B.: Canonpose: Self-supervised monocular 3d human pose estimation in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13294–13304 (2021) 2, 4, 10, 12
51. Wehrbein, T., Rudolph, M., Rosenhahn, B., Wandt, B.: Probabilistic monocular 3d human pose estimation with normalizing flows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11199–11208 (October 2021) 2, 4
52. Xu, H., Bazavan, E.G., Zanfir, A., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: Ghum & ghuml: Generative 3d human shape and articulated pose models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6184–6193 (2020) 4
53. Xu, Y., Wang, W., Liu, T., Liu, X., Xie, J., Zhu, S.C.: Monocular 3d pose estimation via pose grammar and data augmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021) 4
54. Xu, Y., Zhu, S.C., Tung, T.: Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7760–7770 (2019) 4
55. Yu, Z., Ni, B., Xu, J., Wang, J., Zhao, C., Zhang, W.: Towards alleviating the modeling ambiguity of unsupervised monocular 3d human pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8651–8660 (2021) 2, 4, 10, 12
56. Yu, Z., Wang, J., Xu, J., Ni, B., Zhao, C., Wang, M., Zhang, W.: Skeleton2mesh: Kinematics prior injected unsupervised human mesh recovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8619–8629 (2021) 2, 3, 5, 10, 12
57. Zanfir, A., Bazavan, E.G., Zanfir, M., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: Neural descent for visual 3d human pose and shape. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14484–14493 (2021) 4, 10, 11
58. Zanfir, M., Zanfir, A., Bazavan, E.G., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: Thundr: Transformer-based 3d human reconstruction with markers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12971–12980 (October 2021) 4, 10, 11
59. Zeng, W., Ouyang, W., Luo, P., Liu, W., Wang, X.: 3d human mesh regression with dense correspondence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7054–7063 (2020) 4
60. 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) 4, 9
61. Zhang, H., Tian, Y., Zhou, X., Ouyang, W., Liu, Y., Wang, L., Sun, Z.: Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In: Proceedings of the IEEE International Conference on Computer Vision (2021) 11, 14

62. Zheng, M., Planche, B., Gong, X., Yang, F., Chen, T., Wu, Z.: Self-supervised 3d patient modeling with multi-modal attentive fusion. 25th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) (2022) 2
63. Zheng, Z., Yu, T., Wei, Y., Dai, Q., Liu, Y.: Deephuman: 3d human reconstruction from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7739–7749 (2019) 2
64. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5745–5753 (2019) 6
65. Zhu, T., Karlsson, P., Bregler, C.: Simpose: Effectively learning densepose and surface normals of people from simulated data. In: European Conference on Computer Vision. pp. 225–242. Springer (2020) 2
66. Zou, Z., Tang, W.: Modulated graph convolutional network for 3d human pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11477–11487 (October 2021) 4