Self-Supervised 3D Human Mesh Recovery from a Single Image with Uncertainty-Aware Learning

Guoli Yan, Zichun Zhong, Jing Hua

Department of Computer Science, Wayne State University, Detroit, MI, USA
{guoliyan, zichunzhong, jinghua}@wayne.edu

Abstract

Despite achieving impressive improvement in accuracy, most existing monocular 3D human mesh reconstruction methods require large-scale 2D/3D ground-truths for supervision, which limits their applications on unlabeled in-the-wild data that is ubiquitous. To alleviate the reliance on 2D/3D ground-truths, we present a self-supervised 3D human pose and shape reconstruction framework that relies only on self-consistency between intermediate representations of images and projected 2D predictions. Specifically, we extract 2D joints and depth maps from monocular images as proxy inputs, which provides complementary clues to infer accurate 3D human meshes. Furthermore, to reduce the impacts from noisy and ambiguous inputs while better concentrate on the high-quality information, we design an uncertainty-aware module to automatically learn the reliability of the inputs at body-joint level based on the consistency between 2D joints and depth map. Experiments on benchmark datasets show that our approach outperforms other state-of-the-art methods at similar supervision levels.

Introduction

3D human mesh recovery from monocular images is a challenging task in computer vision that can be used for a variety of human-centric applications such as augmented reality, human-robot interaction, computer-assisted coaching, etc. It has received increasing attention in recent years due to the availability of parametric 3D human body model, e.g. SCAPE (Anguelov et al. 2005) and SMPL (Loper et al. 2015), and advances in deep learning techniques (Tian et al. 2023). Although recent monocular 3D human mesh reconstruction methods have gained considerable improvement in accuracy, most of these works are in a fully-supervised setting (Kanazawa et al. 2018; Kolotouros et al. 2019; Lin, Wang, and Liu 2021a,b). Such approaches require large-scale 2D/3D ground truth labels for supervision, restricting their applications on unlabeled in-the-wild data that is abundantly available.

In the absence of 3D ground-truth labels, e.g., SMPL pose and shape parameters, several recent works leverage more easily obtained 2D ground-truth, such as 2D keypoints and silhouette (Pavlakos et al. 2018; Tan, Budvytis, and Cipolla 2017), for a weak supervision. To further alleviate the reliance on paired 2D and 3D ground-truths, attempts are made to regress 3D human pose and shape in a self-supervised manner (Tung et al. 2017a; Kundu et al. 2020; Gong et al. 2022). However, there are still restrictions existing in the previous self-supervised approaches that hinder the generalizability. For example, Gong et al. (2022) requires SMPL data to generate synthetic training data for full supervision, which inherently induces domain gap between synthetic data and real data. Kundu et al. (2020) requires video datasets to generate sequential image pairs for appearance consistency-based self-supervision, which limits its application on datasets with only single-shot images. Different from these methods, we aim to achieve superior generalizability by designing a self-supervised framework that relies only on self-consistency between intermediate representations of images and projected 2D predictions.

Recently, regressing 3D human mesh from intermediate representations (e.g., 2D joints, silhouettes and IUV maps) has achieved promising performance in the self-supervised setting (Tung et al. 2017a; Mugaludi et al. 2021; Gong et al. 2022). These representations can be automatically extracted from RGB images using off-the-shelf algorithms (Cao et al. 2019; Wu et al. 2019; Güler, Neverova, and Kokkinos 2018). Many previous works (Pavlakos et al. 2018; Sengupta, Budvytis, and Cipolla 2020) have explored 2D joints and silhouettes as a combination to provide the pose and shape clues. However, both of them are highly vulnerable to induce pose ambiguity since two different 3D poses may have the same 2D joints and silhouette projection. The depth maps can alleviate such ambiguities and thus can be viewed as a richer substitute to silhouettes. Therefore, we propose to utilize both 2D joints and depth maps that are automatically extracted from images as proxy inputs to infer accurate 3D human meshes.

Although depth maps have been employed in the multi-human reconstruction problem using depth-ordering consistency constraints (Jiang et al. 2020), they are still unexplored for self-supervised 3D human reconstruction. To effectively use depth information, we design depth trimming and depth-point sampling methods for better alignment between input depth and predicted depth. Furthermore, we propose an uncertainty-aware module to automatically learn the reliability of the inputs at body-joint level based on the consistency between 2D joints and depth map. By incorporating these reliability values in the self-consistency losses, the proposed approach can effectively reduce the impacts from noisy and ambiguous inputs while concentrate more on the high-quality information. The contributions of this work can be summarized as follows:

- We propose a simple, novel self-supervised framework that relies only on self-consistency between intermediate representations of images and projected 2D predictions. Without using any 2D/3D ground-truths for supervision, our method can be applied on ubiquitous unlabeled in-the-wild data, achieving superior generalizability.
- We incorporate depth maps in our framework to strengthen the self-consistency constraints, with depth trimming and depth-point sampling designed for better alignment between input depth and predicted depth. To our best knowledge, this work is the first to exploit depth maps for self-supervised 3D human mesh recovery.
- We design an uncertainty-aware module to automatically learn the reliability of the intermediate representations at body-joint level based on the consistency between 2D joints and depth map. The impacts from noisy and ambiguous inputs are effectively reduced by incorporating the reliability values in the self-consistency losses.
- We conduct extensive experiments on benchmark datasets and achieve state-of-the-art results against previous methods at similar supervision levels.

Related Work

3D Human Pose Estimation

The 3D human pose estimation task is commonly formulated as the problem of predicting the 3D positions of body joints from images. Recent approaches can be mainly categorized into image-based and 2D pose-based methods.

The image-based approaches employ the end-to-end learning paradigm, estimating 3D joint locations directly from input images (Pavlakos et al. 2017; Tome, Russell, and Agapito 2017; Zhou et al. 2017; Mehta et al. 2017; Sun et al. 2018; Pavlakos, Zhou, and Daniilidis 2018). For instance, Pavlakos et al. (2017) utilized a volumetric representation for 3D pose and adopted a coarse-to-fine prediction scheme to iteratively refine the 3D joint localization. Sun et al. (2018) proposed an integral regression approach and predicted 3D joint locations in a differentiable way. More recently, Pavlakos, Zhou, and Daniilidis (2018) proposed to use the ordinal depths of human joints as a weak supervision signal to mitigate the need of 3D annotations for 3D pose estimation. However, it is still sub-optimal to train end-to-end 3D pose estimation systems due to the limited availability of 3D captures in the wild and the appearance variations between train and test data.

The 2D pose-based approaches take the intermediately predicted 2D pose as input and lift it to the 3D space (Tung et al. 2017b; Moreno-Noguer 2017; Martinez et al. 2017; Zhao et al. 2019; Wang et al. 2018). For example, Martinez et al. (2017) proposed to use a simple multi-layer perceptron network to regress 3D poses from 2D joint locations. Wang et al. (2018) predicted the depth rankings of body joints by a Pairwise Ranking CNN, and used that as a cue to estimate 3D poses from 2D human joint locations. Zhao et al. (2019) proposed a novel Semantic Graph Convolutional Networks (SemGCN) to capture the spatial relationships between joints for 3D pose regression. These methods gain the advantages of existing 2D pose estimation algorithms to obtain the intermediately estimated 2D poses. Different from the aforementioned methods, our goal is to estimate the whole surface geometry of the human body instead of only 3D joint locations, which is more challenging.

Monocular 3D Human Mesh Recovery

For parametric model-based 3D human pose and shape reconstruction, the goal is to estimate the parameters of the 3D body model, such as SCAPE (Anguelov et al. 2005) and SMPL (Loper et al. 2015). These model-based methods can be further categorized into optimization-based (Bogo et al. 2016; Lassner et al. 2017; Song, Chen, and Hilliges 2020) and regression-based methods (Guler and Kokkinos 2019; Kanazawa et al. 2018; Omran et al. 2018; Choutas et al. 2020; Pavlakos et al. 2018; Tung et al. 2017a).

Optimization-based approaches aim to fit a 3D body model to 2D observations, such as body joints (Bogo et al. 2016) and silhouettes (Lassner et al. 2017). For example, Bogo et al. (2016) proposed a fully automatic approach, SMPLify, to fit the SMPL model to 2D keypoints that are detected by a CNN keypoint detector (Pishchulin et al. 2016). Lassner et al. (2017) extended SMPLify by fitting the SMPL model to body surface landmarks and silhouettes. However, their fitting process is typically very slow and sensitive to initialization.

Regression-based approaches aim to regress the body model parameters from image pixels (Kanazawa et al. 2018; Omran et al. 2018) or intermediate representations such as 2D keypoints and silhouettes (Pavlakos et al. 2018; Tung et al. 2017a). For instance, Kanazawa et al. (2018) proposed HMR to regress SMPL pose and shape parameters directly from image pixels using joint reprojection loss and adversarial prior. Pavlakos et al. (2018) estimated 2D joint heatmaps and the silhouette first before regressing pose parameters from 2D joints and shape parameters from the silhouette. Kolotouros et al. (2019) combined both optimization and regression approaches in one framework. Within a training loop, they used the regressed estimate to initialize SMPLify (Bogo et al. 2016), and used the optimized parameters from SMPLify to supervise the learning of the regressor. More recently, transformer models (Vaswani et al. 2017) have been applied on 3D human mesh recovery domain (Lin, Wang, and Liu 2021a,b), which significantly improve the reconstruction performance.

Self-supervised 3D Human Mesh Recovery

Recent model-based works have also provided self-supervised solutions by leveraging synthetic data (Tung et al. 2017a; Mugaludi et al. 2021; Gong et al. 2022, 2023) or paired appearance consistency (Jiang et al. 2020).

Mugaludi et al. (2021) proposed a self-adaptive approach that uses synthetic data as a source domain to perform full

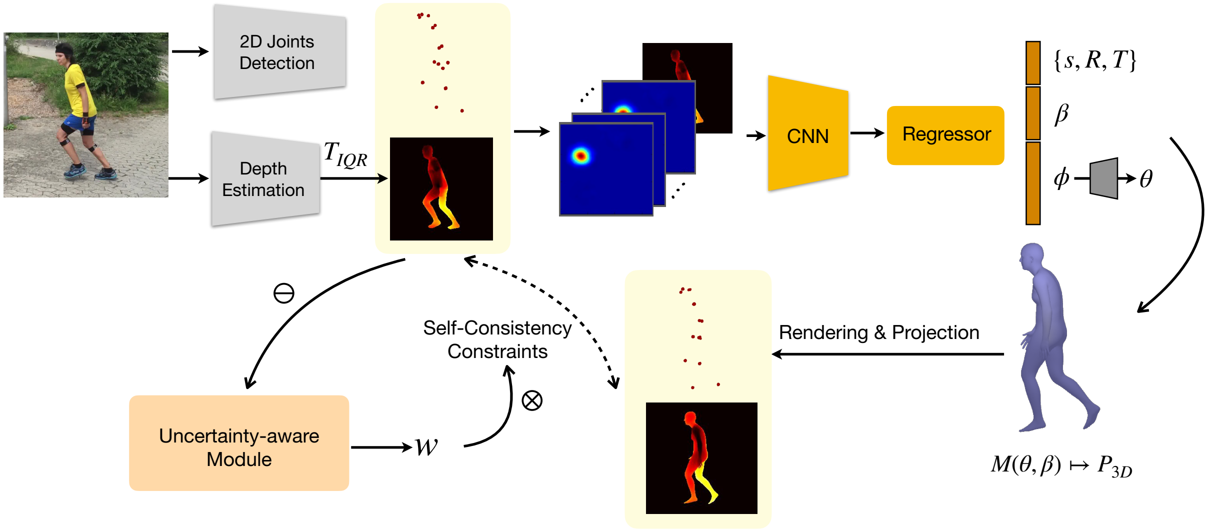

Figure 1: Overview of the proposed approach. The entire framework is trained end-to-end with self-supervision from 2D joints and depth maps. Here $\ominus$ denotes operations to generate relational vectors from the two representations. These relational vectors are the input of the uncertainty-ware module. $\otimes$ means applying the outputs of the uncertainty-ware module as weights to refine the self-consistency constraints.

supervision, and then leverages topological-skeleton that extracted from the raw silhouette to perform self-supervised learning when applying the source-trained model to the unlabeled target domain. Gong et al. (2022) proposed a synthetic-training pipeline that utilizes SMPL data to generate synthetic training data for full supervision, and uses 2D joints and IUV maps as proxy inputs to alleviate the synthetic-to-real gap. Different from these methods, our proposed approach does not require any synthetic data to provide 2D/3D supervision, which inherently bypass the synthetic-to-real gap issue. Kundu et al. (2020) introduced a self-supervised method that relies only on foreground (FG) appearance consistency. This work is close to ours. However, this method requires video datasets to generate sequential image pairs for appearance consensus based self-supervision, which prevents its application to datasets with only single-shot images.

Method

The overall framework of the proposed approach is illustrated in Fig. 1. Given a single image, we use off-the-shelf 2D joints detection algorithms (Cao et al. 2017; Cao et al. 2019; Wu et al. 2019) and depth map estimation methods (Tang et al. 2019; Jafarian and Park 2021) to generate 2D joints and depth maps of humans, respectively. These 2D joints and depth maps serve as the actual inputs of the network and the pseudo-labels to guide the training of the whole network. We also introduce an uncertainty-aware module to automatically learn the reliability of the inputs at body-joint level based on the consistency between 2D joints and depth map. The entire framework is trained end-to-end with self-supervision.

SMPL Human Body Model

Instead of reconstructing 3D human mesh by the network directly, we estimate only a small number of SMPL (Loper et al. 2015) parameters, which are sufficient for generating detailed 3D human mesh by the SMPL model. As a parametric statistical human body model, SMPL represents a 3D human body by Θ, which is composed of pose parameters $\theta \in \mathbb{R}^{72}$ and shape parameters $\beta \in \mathbb{R}^{10}$. The pose parameters contain the relative rotation of 23 joints in axis-angle representation and the global rotation. The shape parameters contain the first 10 coefficients of a PCA shape space. Given the parameters $\Theta = \{\theta, \beta\}$, a triangulated mesh $M(\theta, \beta) \in \mathbb{R}^{3\times N}$ can be generated by the SMPL model, where $N = 6890$ denotes the number of vertices. The major body joints P_{3D} is defined as a linear combination of mesh vertices. Specifically, $P_{3D} = WM$, where W is a pretrained linear regressor.

Proxy Inputs Generation

In the absence of any 2D/3D ground truths, it is still quite challenging to recover 3D human meshes directly from image pixels. To alleviate this issue, we employ intermediate representations of images (i.e., 2D joints and depth maps of humans) as proxy inputs to the regression network, which provides complementary clues to infer 3D human meshes. These representations focus on human bodies, filtering out the information from illumination, background clutter, etc., thus can be viewed as a distillation of RGB images.

2D Joints Generation. With the advances in 2D pose detection approaches in recent years, it is convenient to acquire rather reliable human 2D joints using off-the-shelf methods. Specifically, given an input image, we use Keypoint

R-CNN (He et al. 2017) to predict 2D joint locations. The 2D joints prediction is denoted as $J \in \mathbb{R}^{K\times 2}$, where K is the number of joints. The 2D joints are further transformed into 2D Gaussian joint heatmaps, $F \in \mathbb{R}^{H\times W\times K}$, where H and W represent image height and width, respectively.

Depth Map Trimming. Compared with other intermediate representations such as 2D joints and silhouettes, much less attention has been putted on human depth information for 3D human shape and pose recovery. One main reason is that the estimation of human depth is typically less accurate and less robust compared to other intermediate representations. However, in the self-supervised setting with 2D images only, we find it is beneficial to employ estimated human depth maps for 3D human recovery, which can provide complementary information to 2D joints and alleviate the pose ambiguity issue. Specifically, we employ an unsupervised depth estimation algorithm (Jafarian and Park 2021) to extract depth maps of humans from RGB images. The estimated depth map is denoted as $D \in \mathbb{R}^{H\times W}$.

According to our observations, the original estimated depth maps D are likely to be contaminated by extreme depth values, which results in unreasonable depth ranges for such depth maps, making it hard for the alignment between input depth and predicted depth. To eliminate the impact of outliers (i.e., the extreme depth values), we introduce a depth trimmer T_{IQR} in our framework. Specifically, we apply the interquartile range (IQR) method for calculating the lower and upper bounds of the depth values and trimming off any values that fall outside of the range. Given a depth map D with $H \times W$ depth values, we first get the number of valid (i.e., non-zero) depth values, n, then the IQR is calculated as the difference between the third quartile $Q3$ and the first quartile $Q1$, where $Q1$ is the median of the $\lfloor n/2 \rfloor$ smallest values, and $Q3$ is the median of the $\lfloor n/2 \rfloor$ largest values. The lower and upper bounds are defined as

$$B^{lower} = Q1 - 1.5 \times IQR, \tag{1}$$

$$B^{upper} = Q3 + 1.5 \times IQR. \tag{2}$$

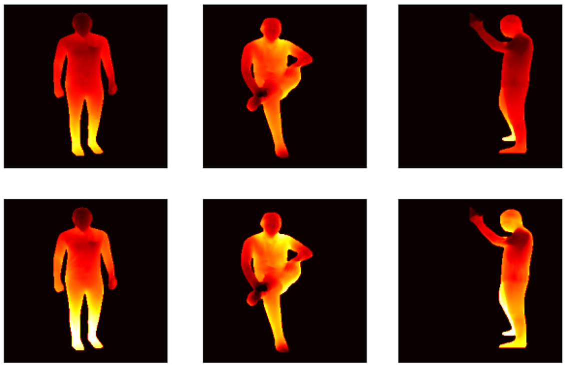

Figure 2: Depth maps before and after IQR trimming. The first row shows original depth maps, the second row shows corresponding trimmed depth maps. All depth maps are normalized for visualization.

Finally, the trimmed depth map $\tilde{D} = T_{IQR}(D)$ is formulated as

$$\tilde{D}_{ij} = \begin{cases} B^{upper}, & D_{ij} > B^{upper}, \\ B^{lower}, & 0 < D_{ij} < B^{lower}, \\ D_{ij}, & otherwise. \end{cases} \tag{3}$$

Fig. 2 illustrates the difference between the original and trimmed depth maps, which are normalized for visualization. It shows that the original depth maps in the 1st row are dominated by dark-color areas due to the existing of outliers, while the trimmed depth maps in the 2nd row have more balanced dark-to-bright color distribution, which are more faithful to the ground truth.

Uncertainty Modeling

Due to the differences in view point, occlusion condition, pose topology, etc., the quality of the pre-extracted 2D joints and depth maps usually varies among different image samples as well as different human parts within a sample. In our self-supervised framework, we take the above intermediate representations as pseudo ground truth labels to guide the learning of the whole network. Therefore, it is critical to quantify the uncertainty of each pseudo ground-truth, so that the reliable ones can be better concentrated while the impact of noisy and ambiguous ones can be alleviated.

We define the uncertainty at body-joint level based on the consistency between 2D joints and depth map from the same image. The measure of consistency between the two representations is based on the symmetrical nature of human bodies, e.g., the length of left upper-arm is equal to the length of right upper-arm typically. Thus if the depth difference of left upper-arm is close to that of right upper-arm, their bone lengths should be close in the 2D skeleton that derived from 2D joints. Specifically, we denote each bone in the left body as B_i^l, and the symmetrically corresponding bone in the right body as B_i^r, where i is a shared bone-index among left and right body part. Since each bone is a connection between two joints, we define the 2D bone length $Len(\cdot)$ as the Euclidean distance between the connected two joints, and the bone depth discrepancy $DD(\cdot)$ as the depth difference between the two joints. According to the bone symmetricity and 3D-to-2D projection properties, we have the following observations: (1) the smaller the $DD(\cdot)$, the larger the $Len(\cdot)$; (2) The closer between $DD(B_i^l)$ and $DD(B_i^r)$, the closer between $Len(B_i^l)$ and $Len(B_i^r)$. The conformity of such relationships among bone lengths and depth discrepancies reflects the consistency between 2D joints and depth map. Therefore, we can learn the uncertainty of proxy inputs using the bone lengths and depth discrepancies. We utilize a multilayer perceptron (MLP) network that contains two fully connected layers and a Sigmoid function to automatically learn joint-level reliability:

$$w = Sigmoid(FC(FC(v_{Len} \oplus v_{DD}))), \tag{4}$$

where $\oplus$ means concatenation operation, v_{Len} is a vector of bone lengths calculated by $Len(\cdot)$, and v_{DD} is a vector of depth discrepancies calculated by $DD(\cdot)$.

Self-Supervised Human Reconstruction

Pose Prior. In order to prevent the network from producing physically implausible 3D bodies, we employ a pose prior model in our framework similar to (Kundu et al. 2020; Jafarian and Park 2021). The pose prior is the decoder of an adversarial auto-encoder (Makhzani et al. 2015) trained on a large amount of 3D pose samples (Mahmood et al. 2019). It learns a latent pose feature $\phi \in [-1, 1]^{32}$ in the bottleneck, and the decoder is learned to recover the realistic SMPL pose $\theta \in \mathbb{R}^{69}$ of 23 joints from ϕ. We only take the decoder from the trained auto-encoder as the pose prior model in our framework, and keep its weights frozen during our self-supervised training.

Network Architecture. Similar to HMR (Kanazawa et al. 2018), we use the ResNet-18 (He et al. 2016) as the CNN backbone. The input is the concatenation of joint heatmaps F and trimmed depth map $\tilde{D}$ along the channel dimension, which results in a tensor of shape $H \times W \times (K + 1)$. The output of ResNet is average pooled, which produces features $f \in \mathbb{R}^{512}$. The subsequent regression module consists of two fully connected layers with 512 neurons each, followed by an output layer with 48 neurons, which contains the camera parameters $\{s, R, T\}$, where $s \in \mathbb{R}$ and $T \in \mathbb{R}^2$ denotes the scale and translation, respectively, $R \in \mathbb{R}^3$ is the global rotation. The output layer also contains the SMPL shape parameter $\beta \in \mathbb{R}^{10}$ and a pose embedding vector $\phi \in \mathbb{R}^{32}$, which is then sent to the pose prior to generate the SMPL pose parameter $\theta \in \mathbb{R}^{69}$ of 23 joints.

Self-Consistency Loss. To conduct self-consistency on 2D joints, the estimated SMPL parameters $\Theta = \{\theta, \beta\}$ are transformed into 2D joints J' through 3D joints regression from reconstructed mesh and weak-perspective projection using the estimated camera parameters. Then the self-consistency loss for 2D joints can be expressed as

$$L_{joint}(J, J') = \sum_{i=1}^{K} \left\| J_i - J_i' \right\|_2^2 . \tag{5}$$

For self-consistency on depth maps, the estimated SMPL parameters are transformed into depth map D' through differentiable rendering (Kato, Ushiku, and Harada 2018) and weak-perspective projection. However, we do not calculate the L_2 distance between $\tilde{D}$ and D' directly. Instead, we evenly sample depth points on bones in order to obtain better correspondences between $\tilde{D}$ and D'. Specifically, we keep the two depth points on joints and sample the remaining ones evenly along each bone. The depth loss is calculated between each depth point $\tilde{D}_p^i$ and its corresponding point $D_p^{i\prime}$, i.e.,

$$L_{depth}(\tilde{D}, D') = \sum_{i=1}^{m} \left\| \tilde{D}_p^i - D_p^{i\prime} \right\|_2^2 , \tag{6}$$

where m is the total number of depth points. The reliability of each depth point is correlated with the the reliability of the two joints on the same bone. For simplicity, we define the reliability w_{p_i} of each depth point p_i as the reliability of the joint with closer distance. We take the depth point reliability as a weight to boost network training. Then the uncertainty-aware depth consistency loss is defined as

$$L^{*}_{depth}(\tilde{D}, D') = \sum_{i=1}^{m} w_{p_i} \left\| \tilde{D}_p^i - D_p^{i\prime} \right\|_2^2 . \tag{7}$$

Our final loss function is defined as

$$L = \alpha_j L_{joint}(J, J') + \alpha_d L^{*}_{depth}(\tilde{D}, D'), \tag{8}$$

where α_j and α_d are loss weights for the joint and depth, respectively.

Experiments

Datasets

In our experiments, we use Human3.6M (Ionescu et al. 2013), 3DPW (Von Marcard et al. 2018) and UP-3D (Lassner et al. 2017) for training. For 3DPW (Von Marcard et al. 2018) and Human3.6M (Ionescu et al. 2013), we report evaluation results using mean per joint position error (MPJPE) and Procrustes-aligned mean per joint position error (PA-MPJPE). Note that we only use image data and no 2D/3D annotations from these datasets are involved during training. More detailed information of these datasets is provided in the following.

Human3.6M is a large-scale indoor dataset captured in a controlled environment. Its training set contains 7 subjects performing 4 types of actions under 4 camera views. Following the Protocol 2 (Kanazawa et al. 2018), we train our model on 5 subjects (S1, S5, S6, S7, S8) and test on the front-view samples of the rest 2 subjects (S9, S11). All videos are downsampled from 50fps to10fps.

3DPW is an in-the-wild dataset that contains both indoor and outdoor scenes. The dataset has 60 video sequences. Both training and testing sets contain 24 videos, and the rest 12 video are used for validation. Following (Kocabas, Athanasiou, and Black 2020), we use its training data when conducting experiments on 3DPW.

UP-3D is an outdoor dataset. It contains more than 8K images. This dataset is only used for training.

Implementation Details

The MLP network for uncertainty modeling contains two fully connected layers with 128 and 64 neurons, respectively. The output layer contains 12 neurons, which is equal to the number of joints excluding joints on the head. We set the joint loss weight $\alpha_j = 1$ and the depth loss weight $\alpha_d = 0.04$. We use the Adam optimizer (Kingma and Ba 2014) with an initial learning rate of 10^{-5}, and batch size of 64. After training for 10 epochs, we regularize β to remain close to the mean shape. Our experiments run on a single NVIDIA GeForce RTX 3090 GPU.

Ablation Study

In Table 1, we compare our proposed model with several variants on 3DPW dataset to investigate the contribution of each component. Specifically, we design the following baselines: (1) "Ours $\ominus$ uncertainty" denotes our model using

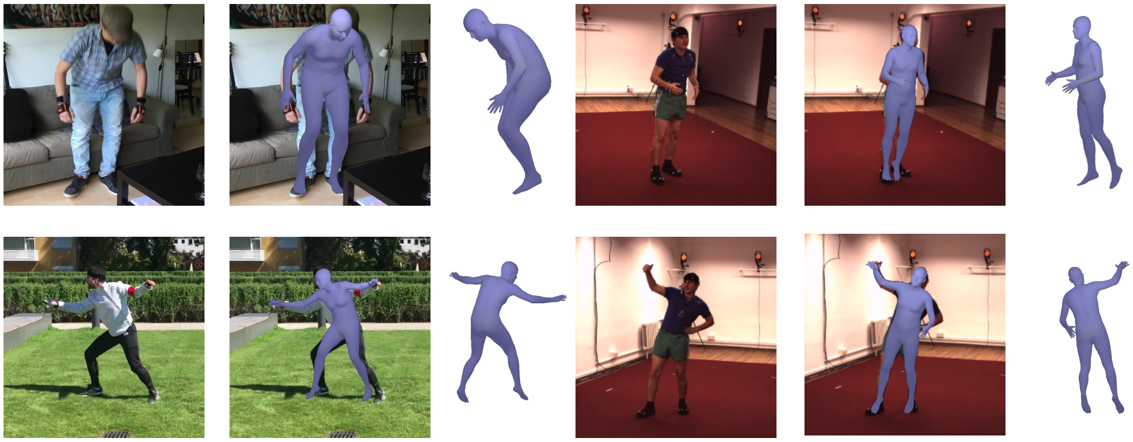

Figure 3: Examples of qualitative results on two datasets. Left three columns: 3DPW dataset. Right three columns: Human3.6M dataset. The 3rd and 6th columns show another view of the 3D reconstruction results.

Methods	MPJPE	PA-MPJPE
Ours	159.4	89.5
Ours $\ominus$ uncertainty	166.9	93.4
Ours $\ominus$ uncertainty $\ominus$ T_{IQR}	208.2	113.5
Ours $\ominus$ uncertainty $\ominus$ sampling	182.6	99.2
Ours $\ominus$ depth input	165.9	97.6

Table 1: Ablation study on 3DPW dataset. The MPJPE and PA-MPJPE (both in mm) are reported.

Sup.	Methods	MPJPE	PA-MPJPE
Full	HMR (2018)	128.1	81.3
	SPIN (2019)	98.6	59.2
	PyMAF (2021)	92.8	58.9
Weak	SMPLify (2016)	199.2	106.1
	Mugaludi et al. (2021) (S→R, weak)	126.3	79.1
Self-sup. (use syn.)	RGB Only (2019)	-	105.6
	Flow Only (2019)	-	100.1
	Mugaludi et al. (2021)	159.0	95.1
Self-sup.	Kundu et al. (2020)	187.1	102.7
	Ours	**159.4**	**89.5**

Table 2: Comparison with the state-of-the-art methods on 3DPW in terms of MPJPE and PA-MPJPE (both in mm).

L_{depth} for depth loss instead of using uncertainty-weighted depth loss L^{*}_{depth}; (2) "Ours $\ominus$ uncertainty $\ominus$ T_{IQR}" denotes "Ours $\ominus$ uncertainty" using the original depth maps as pseudo ground truths instead of using IQR trimmed depth maps; (3) "Ours $\ominus$ uncertainty $\ominus$ sampling" denotes our method self-supervised on depth maps directly without depth points sampling and uncertainty awareness; (4) "Ours $\ominus$ depth input" denotes our model using only 2D joints as input while keep the same loss functions. All these models are trained on the training images from 3DPW and UP3D.

Compared with our proposed model, the reconstruction error is increased by 7.5 (MPJPE) and 3.9 (PA-MPJPE) on the baseline "Ours $\ominus$ uncertainty awareness". This shows the effectiveness of using uncertainty-aware depth loss, which boosts the self-supervised learning more effectively while alleviates the impacts from high-uncertainty depth. The performance of "Ours $\ominus$ uncertainty awareness $\ominus$ T_{IQR}" is further degraded by a large margin, which demonstrates the effectiveness of applying IQR trimming on the depth map. Without depth trimming, the outliers will cause unreasonable depth range, making it hard for the alignment between D and D' along depth dimension. Actually, its results are even worse than not using depth for self-supervision. Compared with our model, the MPJPE and PA-MPJPE are increased by 6.5 and 8.1 on the baseline "Ours $\ominus$ depth input", which shows the complementary effect of depth to the 2D joints. In fact, it is highly possible to have pose ambiguity using 2D joints only since two different 3D poses may have same 2D projection. The depth map can alleviate such ambiguities.

Comparison with the State-of-the-Art

We compare the reconstruction performance of our method with previous state-of-the-art methods of different supervision degrees on 3DPW and Human3.6M datasets. The results are shown in Table 2 and Table 3.

On 3DPW dataset (Table 2), our method achieves the state-of-the-art performance among the self-supervised methods. Specifically, our method outperforms Kundu et al. (2020) by a large margin. The reconstruction error of our method is decreased by 27.7 (MPJPE) and 13.2 (PA-MPJPE). Our method also outperforms Flow Only (Doersch and Zisserman 2019) and Mugaludi et al. (2021) in terms of PA-MPJPE. Note that Mugaludi et al. (2021) is fully supervised on the synthetic data (source domain) and adapt to a target domain with self-adaption, while we do not need any synthetic data for 2D/3D supervision.

Sup.	Methods	PA-MPJPE
Full	Lassner et al. (2017)	93.9
	Pavlakos et al. (2018)	75.9
	HMR (2021a)	56.8
	Kolotouros et al. (2019)	50.1
	SPIN (2019)	41.1
Weak	HMR (unpaired) (2021a)	66.5
	SPIN (unpaired) (2019)	62.0
	Mugaludi et al. (2021) (S→R, weak)	58.1
Self-sup. (use syn.)	Tung et al. (2017a)	98.4
	Mugaludi et al. (2021)	81.3
Self-sup.	Rhodin et al. (2018)	98.2
	Kundu et al. (2020)	90.5
	Ours	**85.4**

Table 3: Comparison with the state-of-the-art methods on Human3.6M dataset using Protocol 2.

On Human3.6 dataset (Table 3), our method consistently outperforms Kundu et al. (2020), which is directly comparable with our method due to the similar supervision setting. The reconstruction error of our method is decreased by 5.1 (PA-MPJPE) compared with Kundu et al. (2020), even though we do not require any paired images for training. In addition, our result is also comparable with Mugaludi et al. (2021), which requires synthetic data for full supervision, while we do not have such requirements.

Qualitative Results

We have also evaluated our models qualitatively on 3DPW and Human3.6 datasets. Some examples of our results are presented in Fig. 3. We observe that the postures of humans are well captured with our proposed method, although we do not use any 2D/3D ground-truths from these dataset for training. By incorporating the estimated depth maps in our framework, the pose ambiguity issue is relatively alleviated.

In Fig. 4, we qualitatively compare the performance of our method with baseline models on the 3DPW dataset. It shows that the reconstruction results of our method are consistently better than all the baseline models. In addition, the performance of baseline "Ours $\ominus$ uncertainty" is also better than the other two baselines qualitatively, which is in line with the quantitative results in Table 1. These further demonstrate the effectiveness of each proposed component.

Limitations. In the circumstance of severe occlusion, it is quite challenging to get accurate intermediate representations from images, thus our method tends to fail in such cases (see Fig. 5). From the qualitative results in Fig. 3, we also observe that the reconstruction on feet and hands is still not so desirable compared with existing supervised methods. The main reason is that we use a sparse 2D pose representation which has no joint on feet or hands other than the ankle or wrist joints. This issue can be alleviated by adding more extra joints on the region of interest to enhance the accuracy based on the task need.

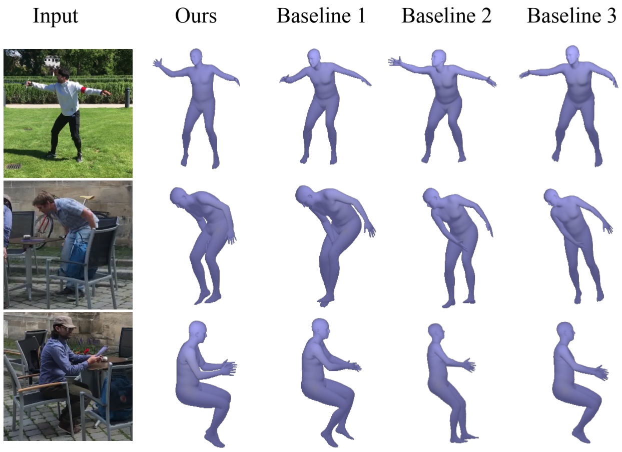

Figure 4: Qualitative comparison with baseline models on 3DPW. Our results are in the 2nd column. The following columns show results of baseline "Ours $\ominus$ uncertainty", "Ours $\ominus$ uncertainty $\ominus$ T_{IQR}", and "Ours $\ominus$ uncertainty $\ominus$ sampling", respectively.

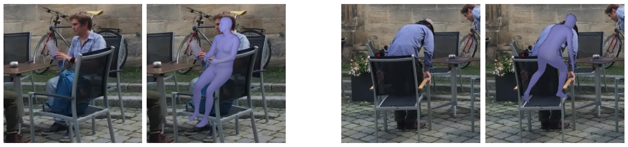

Figure 5: Failure cases caused by severe occlusion.

Conclusion

In this work, we have presented a simple, novel self-supervised framework for 3D human mesh recovery from monocular images with uncertainty-aware learning. The proposed method does not require any 2D/3D ground-truths for supervision, relying only on self-consistency between intermediate representations, i.e., 2D joints and depth maps, and projected ones after 3D human reconstruction. We incorporate depth maps in our framework to strengthen the self-consistency constraints, with depth trimming and depth-point sampling methods designed for better alignment between input depth and predicted depth. Furthermore, we design an uncertainty-aware module to automatically learn the reliability of the intermediate representations at body-joint level based on the consistency between two representations. The impacts from noisy and ambiguous inputs are effectively reduced by incorporating the reliability values in the self-consistency losses. Experiments demonstrate the effectiveness of our proposed method.

Acknowledgements

We would like to thank the reviewers for their valuable comments. This work was partially supported by the NSFC 61972353 and the NSF under Grant Numbers IIS-1816511, OAC-1845962, OAC-1910469, and OAC-2311245.

References

Anguelov, D.; Srinivasan, P.; Koller, D.; Thrun, S.; Rodgers, J.; and Davis, J. 2005. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, 408–416.

Bogo, F.; Kanazawa, A.; Lassner, C.; Gehler, P.; Romero, J.; and Black, M. J. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European conference on computer vision*, 561–578. Springer.

Cao, Z.; Hidalgo Martinez, G.; Simon, T.; Wei, S.; and Sheikh, Y. A. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR*.

Choutas, V.; Pavlakos, G.; Bolkart, T.; Tzionas, D.; and Black, M. J. 2020. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision*, 20–40. Springer.

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

Gong, X.; Song, L.; Zheng, M.; Planche, B.; Chen, T.; Yuan, J.; Doermann, D.; and Wu, Z. 2023. Progressive Multi-View Human Mesh Recovery with Self-Supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 676–684.

Gong, X.; Zheng, M.; Planche, B.; Karanam, S.; Chen, T.; Doermann, D.; and Wu, Z. 2022. Self-supervised Human Mesh Recovery with Cross-Representation Alignment. In *European Conference on Computer Vision*, 212–230. Springer.

Guler, R. A.; and Kokkinos, I. 2019. Holopose: Holistic 3D human reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10884–10894.

Güler, R. A.; Neverova, N.; and Kokkinos, I. 2018. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7297–7306.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2013. 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.

Jafarian, Y.; and Park, H. S. 2021. Learning high fidelity depths of dressed humans by watching social media dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12753–12762.

Jiang, W.; Kolotouros, N.; Pavlakos, G.; Zhou, X.; and Daniilidis, K. 2020. Coherent reconstruction of multiple humans from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5579–5588.

Kanazawa, A.; Black, M. J.; Jacobs, D. W.; and Malik, J. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7122–7131.

Kato, H.; Ushiku, Y.; and Harada, T. 2018. Neural 3D mesh renderer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3907–3916.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kocabas, M.; Athanasiou, N.; and Black, M. J. 2020. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5253–5263.

Kolotouros, N.; Pavlakos, G.; Black, M. J.; and Daniilidis, K. 2019. 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*, 2252–2261.

Kolotouros, N.; Pavlakos, G.; and Daniilidis, K. 2019. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4501–4510.

Kundu, J. N.; Rakesh, M.; Jampani, V.; Venkatesh, R. M.; and Venkatesh Babu, R. 2020. Appearance consensus driven self-supervised human mesh recovery. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, 794–812. Springer.

Lassner, C.; Romero, J.; Kiefel, M.; Bogo, F.; Black, M. J.; and Gehler, P. V. 2017. Unite the people: Closing the loop between 3D and 2D human representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6050–6059.

Lin, K.; Wang, L.; and Liu, Z. 2021a. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1954–1963.

Lin, K.; Wang, L.; and Liu, Z. 2021b. Mesh graphormer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12939–12948.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6): 1–16.

Mahmood, N.; Ghorbani, N.; Troje, N. F.; Pons-Moll, G.; and Black, M. J. 2019. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5442–5451.

Makhzani, A.; Shlens, J.; Jaitly, N.; Goodfellow, I.; and Frey, B. 2015. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*.

Martinez, J.; Hossain, R.; Romero, J.; and Little, J. J. 2017. A simple yet effective baseline for 3D human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2640–2649.

Mehta, D.; Rhodin, H.; Casas, D.; Fua, P.; Sotnychenko, O.; Xu, W.; and Theobalt, C. 2017. Monocular 3D human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, 506–516. IEEE.

Moreno-Noguer, F. 2017. 3D human pose estimation from a single image via distance matrix regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2823–2832.

Mugaludi, R. R.; Kundu, J. N.; Jampani, V.; et al. 2021. Aligning silhouette topology for self-adaptive 3D human pose recovery. *Advances in Neural Information Processing Systems*, 34: 4582–4593.

Omran, M.; Lassner, C.; Pons-Moll, G.; Gehler, P.; and Schiele, B. 2018. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 international conference on 3D vision (3DV)*, 484–494. IEEE.

Pavlakos, G.; Zhou, X.; and Daniilidis, K. 2018. Ordinal depth supervision for 3D human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7307–7316.

Pavlakos, G.; Zhou, X.; Derpanis, K. G.; and Daniilidis, K. 2017. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7025–7034.

Pavlakos, G.; Zhu, L.; Zhou, X.; and Daniilidis, K. 2018. 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*, 459–468.

Pishchulin, L.; Insafutdinov, E.; Tang, S.; Andres, B.; Andriluka, M.; Gehler, P. V.; and Schiele, B. 2016. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4929–4937.

Rhodin, H.; Salzmann, M.; and Fua, P. 2018. Unsupervised geometry-aware representation for 3d human pose estimation. In *Proceedings of the European conference on computer vision (ECCV)*, 750–767.

Sengupta, A.; Budvytis, I.; and Cipolla, R. 2020. Synthetic training for accurate 3D human pose and shape estimation in the wild. *arXiv preprint arXiv:2009.10013*.

Song, J.; Chen, X.; and Hilliges, O. 2020. 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*, 744–760. Springer.

Sun, X.; Xiao, B.; Wei, F.; Liang, S.; and Wei, Y. 2018. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 529–545.

Tan, J. K. V.; Budvytis, I.; and Cipolla, R. 2017. Indirect deep structured learning for 3D human body shape and pose prediction. *BMVC*.

Tang, S.; Tan, F.; Cheng, K.; Li, Z.; Zhu, S.; and Tan, P. 2019. A neural network for detailed human depth estimation from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7750–7759.

Tian, Y.; Zhang, H.; Liu, Y.; and Wang, L. 2023. Recovering 3d human mesh from monocular images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Tome, D.; Russell, C.; and Agapito, L. 2017. Lifting from the deep: Convolutional 3D pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2500–2509.

Tung, H.-Y.; Tung, H.-W.; Yumer, E.; and Fragkiadaki, K. 2017a. Self-supervised learning of motion capture. *Advances in neural information processing systems*, 30.

Tung, H.-Y. F.; Harley, A. W.; Seto, W.; and Fragkiadaki, K. 2017b. Adversarial inverse graphics networks: Learning 2D-to-3D lifting and image-to-image translation from unpaired supervision. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 4364–4372. IEEE.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Von Marcard, T.; Henschel, R.; Black, M. J.; Rosenhahn, B.; and Pons-Moll, G. 2018. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, 601–617.

Wang, M.; Chen, X.; Liu, W.; Qian, C.; Lin, L.; and Ma, L. 2018. Drpose3D: Depth ranking in 3D human pose estimation. *arXiv preprint arXiv:1805.08973*.

Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.-Y.; and Girshick, R. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

Zhang, H.; Tian, Y.; Zhou, X.; Ouyang, W.; Liu, Y.; Wang, L.; and Sun, Z. 2021. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11446–11456.

Zhao, L.; Peng, X.; Tian, Y.; Kapadia, M.; and Metaxas, D. N. 2019. Semantic graph convolutional networks for 3D human pose regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3425–3435.

Zhou, X.; Huang, Q.; Sun, X.; Xue, X.; and Wei, Y. 2017. Towards 3D human pose estimation in the wild: a weakly-supervised approach. In *Proceedings of the IEEE International Conference on Computer Vision*, 398–407.