Implicit 3D Human Mesh Recovery using Consistency with Pose and Shape from Unseen-view

Hanbyel Cho Yooshin Cho Jaesung Ahn Junmo Kim
Korea Advanced Institute of Science and Technology (KAIST), South Korea
{tlrl4658,choys95,jaesung02,junmo.kim}@kaist.ac.kr

Abstract

*From an image of a person, we can easily infer the natural 3D pose and shape of the person even if ambiguity exists. This is because we have a mental model that allows us to imagine a person's appearance at different viewing directions from a given image and utilize the consistency between them for inference. However, existing human mesh recovery methods only consider the direction in which the image was taken due to their structural limitations. Hence, we propose "**Imp**licit 3D **H**uman **M**esh **R**ecovery (**ImpHMR**)" that can implicitly imagine a person in 3D space at the feature-level via Neural Feature Fields. In ImpHMR, feature fields are generated by CNN-based image encoder for a given image. Then, the 2D feature map is volume-rendered from the feature field for a given viewing direction, and the pose and shape parameters are regressed from the feature. To utilize consistency with pose and shape from unseen-view, if there are 3D labels, the model predicts results including the silhouette from an arbitrary direction and makes it equal to the rotated ground-truth. In the case of only 2D labels, we perform self-supervised learning through the constraint that the pose and shape parameters inferred from different directions should be the same. Extensive evaluations show the efficacy of the proposed method.*

1. Introduction

Human Mesh Recovery (HMR) is a task that regresses the parameters of a three-dimensional (3D) human body model (*e.g.*, SMPL [34], SMPL-X [42], and GHUM [57]) from RGB images. Along with 3D joint-based methods [7, 32, 46], HMR has many downstream tasks such as AR/VR, and computer graphics as a fundamental topic in computer vision. In recent years, there has been rapid progress in HMR, particularly in regression-based approaches [6, 19, 22, 25–27, 30, 49, 55, 62]. However, despite these achievements, the existing algorithms still have a gap with the way humans do, so most of them do not show robust performance against the inherent ambiguity of the task.

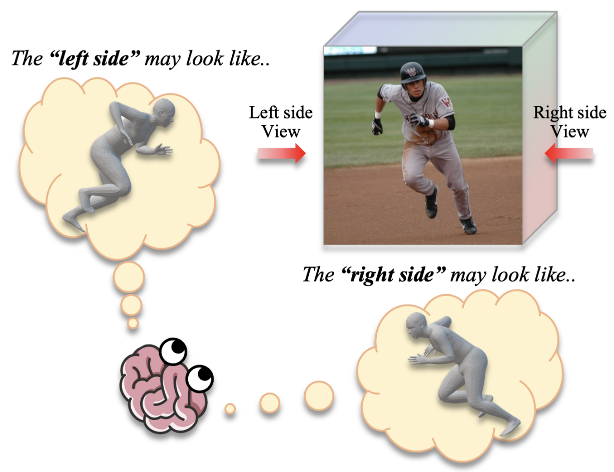

Figure 1. **Mental model of human that infers pose and shape from a single image.** From an image of a person, we infer pose and shape robustly by imagining the person's appearance not only from the direction in which the image was taken, but also from other viewing directions (*e.g.*, left and right sides).

Consider the image of a baseball player running, as shown in Fig. 1. For the given single image, we can easily infer that the person's right elbow and left leg are extended backward in a 3D space, despite the presence of inherent ambiguity (*e.g.*, depth and occlusion). This is because we have a mental model that allows us to imagine a person's appearance at different viewing directions from a given image and utilize the consistency between them for inference. Recently, many state-of-the-art studies have successfully utilized knowledge similar to that used by humans such as human dynamics [20] and temporal information [8, 23, 35, 55]. However, to the best of our knowledge, there have been no studies proposed methods that consider 3D space for HMR similar to the way we infer pose and shape through appearance check between different views in 3D space.

To overcome this issue, we propose "Implicit 3D Human Mesh Recovery (ImpHMR)" that can implicitly imagine a human placed in a 3D space via Neural Feature Fields [40]. Our assumption is that if the model is trained to infer a human's pose and shape at arbitrary viewing directions in a 3D space from a single image, then the model learns better spatial prior knowledge about human appearance; consequently, the performance in the canonical viewing direction in which the image was taken is improved.

To achieve this, we incorporate Neural Feature Fields into regression-based HMR methods. In ImpHMR, it generates feature fields using a CNN-based image encoder for a given image to construct a person in 3D space, as shown in Fig. 2. A feature field represented by a Multi-Layer Perceptron (MLP) is a continuous function that maps the position of a point in 3D space and a ray direction to a feature vector and volume density. In a feature field, which is implicit representation, all continuous points in a space can have a respective feature and volume density. Hence, the feature field is more expressive than explicit representation [59] and more suitable for representing human appearance from different viewing directions in 3D space.

To infer the pose and shape parameters from the Feature Field, the 2D feature map is generated by volume rendering for a given viewing direction, and the parameters are regressed from the rendered feature. Unlike previous methods, our model can look at a person from an *arbitrary viewing direction* by controlling the viewing direction determined by camera extrinsic (*i.e*., camera pose). Therefore, to utilize consistency with pose and shape from unseen-view, if there are 3D labels, ImpHMR predicts results including silhouette used as *geometric guidance* from an arbitrary direction and makes it equal to the rotated ground-truth. In addition, in the case of only 2D labels, we perform self-supervised learning through the constraint that SMPL parameters inferred from different directions should be the same. These constraints help feature fields represent a better 3D space by disentangling human appearance and viewing direction; as a result, SMPL regression from canonical viewing direction in which the image was taken is improved. To verify the efficacy of our method, we conduct experiments on 3DPW, LSP, COCO, and 3DPW-OCC. The contributions of our work can be summarized as follows:

- We propose a novel HMR model called "ImpHMR" that can implicitly imagine a human in 3D space from a given 2D observation via Neural Feature Fields.
- To utilize consistency with pose and shape from unseen-view, we propose arbitrary view imagination loss and appearance consistency loss.
- We propose the geometric guidance branch so that the model can learn better geometric information.
- ImpHMR has $2 \sim 3$ *times faster* fps than current SOTAs thanks to efficient spatial representation in feature fields.
- We confirm that having the model imagine a person in 3D space and checking consistency between human appearance from different viewing directions improves the HMR performance in the canonical viewing direction in which the image was taken.

2. Related Work

2.1. Human Mesh Recovery

Human mesh recovery works have been conducted based on two approaches: *optimization-based* approaches [3, 29] and *regression-based* approaches [19, 41, 45]. Recent works tend to focus on regression-based approaches.

Optimization-based Approaches. Early works in this field have mainly focused on the optimization-based approaches fitting parametric human body models. SMPLify [3] fits the parametric model, SMPL [34] to minimize errors between the projection of recovered meshes and 2D/3D evidence, such as silhouettes or keypoints. In addition, prior terms are adopted to penalize the unrealistic shape and pose. In subsequent studies, 2D/3D information was utilized in the fitting procedure, and optimization with more expressive models in the multi-view has been suggested [18, 29, 61, 64]. Recently, hybrid approach, which combining optimization and regression-based approaches, has been proposed and provides a more accurate pseudo ground-truth 3D (*e.g*., SPIN [25] and EFT [17]) for 2D images. Despite the accurate results generated via optimization-based approaches, the fitting processes still remained slow and sensitive to initialization.

Regression-based Approaches. To avoid the issues of optimization-based methods, recent works have adopted regression-based approaches and utilized the powerful learning capability of deep neural networks [6, 9, 10, 14, 19, 25, 41, 45]. Deep networks were directly used to regress model parameters from a single RGB image and supervised with 2D/3D annotations, such as 3D shape ground truth, keypoints, silhouettes, and parts segmentation. Regression-based methods have made significant advances by adopting network architectures that were suitable to learn different types of supervision signals [28, 39, 43–45, 47, 52, 53, 58, 60]. Zhang *et al*. [62] proposed a pyramidal mesh alignment feedback that allows images and meshes to be well aligned, paying attention to the fact that there is no forward feedback when conventional regressors infer SMPL parameters iteratively. In addition, Li *et al*. [30] proposed a hybrid approach with joint estimation using inverse kinematics, and [24, 63] and [22, 49] proposed a method for a situation with occlusion and multi-person, respectively. Furthermore, recent studies have successfully utilized knowledge similar to that used by humans, such as human dynamics [20] and temporal information [8, 23, 35, 55]. However, to the best of our knowledge, there have been no studies that proposed methods that consider 3D space for HMR similar to the way we infer pose and shape through appearance checks between different views in 3D space.

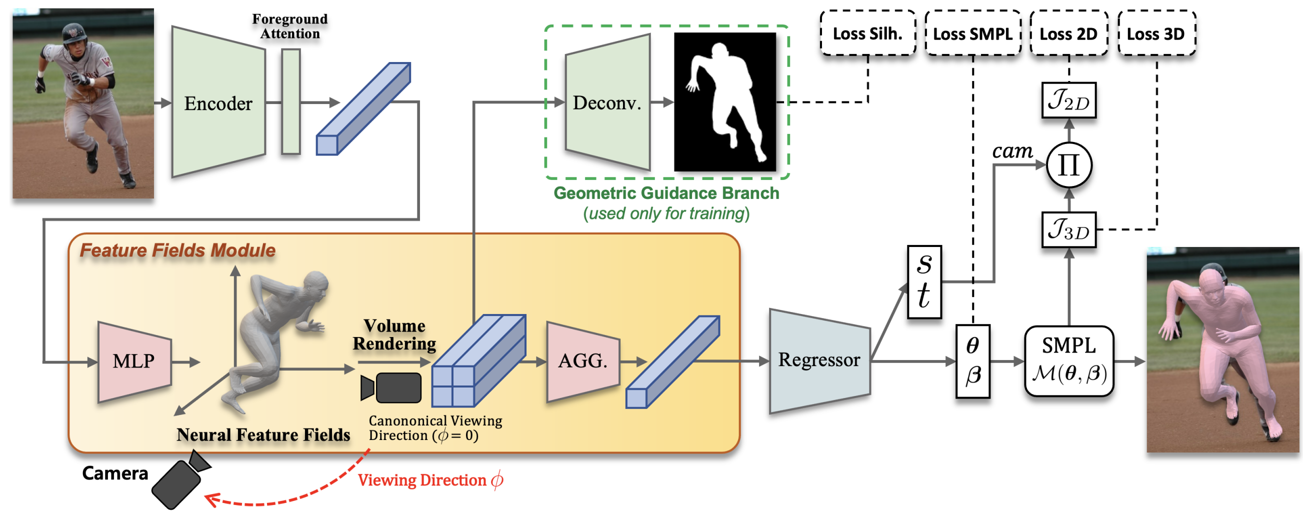

Figure 2. **Overview of ImpHMR architecture.** Given an image of a person, ImpHMR can implicitly imagine the person in 3D space and infer SMPL parameters viewed from an arbitrary viewing direction ϕ through *Feature Fields Module*. The model infers parameters from arbitrary directions during training to have a better 3D prior about person; consequently, regression performance in *Canonical Viewing Direction* is improved. For simplicity, we omit notation ϕ and write loss functions in Sec 3.4 abstractly according to the form of the output.

2.2. Implicit Neural Representations

Neural Radiance Fields. Previously, in 3D reconstruction, differentiable rendering techniques have been adopted to overcome the requirement for 3D supervision [4, 33]. A radiance field is a continuous function whose input is a set of a 3D location and a 2D ray direction, and its output is an RGB color value and a volume density [5, 37]. To exploit the effective non-linear mapping capability of deep neural networks, Mildenhall *et al.* [38] proposed to learn Neural Radiance Fields (NeRFs) by parameterizing with a Multi-Layer Perceptron (MLP) and successfully combining with volume rendering for novel view synthesis.

Neural Feature Fields. Since the success of NeRFs [38], generative models for neural radiance fields have been proposed [40, 48]. To get better representations of objects, Niemeyer *et al.* [40] proposed Generative Neural Feature Fields (GIRAFFE) that replace the color output with a generic feature vector. In addition, neural feature fields condition the MLP on latent vectors of the shape and appearance of objects. Therefore, unlike NeRFs that fits the MLP to multi-view images of a single scene, neural feature fields have the capability to generate novel scenes. In this study, we adopt Neural Feature Fields to design a mental model that imagines a human in a 3D space from a single image.

3. Methodology

The overall framework of the proposed method is shown in Fig. 2. In this section, we provide a detailed explanation of the proposed method. First, we recapitulate the outline of Neural Feature Fields [40] and SMPL body model [34]. Then, we describe the model architecture and training objective of the proposed method.

3.1. Neural Feature Fields

A Neural Feature Field [40] is a continuous function h that maps a 3D point $\mathbf{x} \in \mathbb{R}^3$ and a ray direction $\mathbf{r} \in \mathbb{S}^2$ to a volume density $\sigma \in \mathbb{R}^+$ and an M_f-dimensional feature vector $\mathbf{f} \in \mathbb{R}^{M_f}$. When h is parameterized by a deep neural network, the low-dimensional input $\mathbf{x}$ and $\mathbf{r}$ are first mapped to higher-dimensional features through the positional encoding [38, 51] so that they can be mapped to feature vectors $\mathbf{f}$ capable of representing complex scenes. Concretely, each element of $\mathbf{x}$ and $\mathbf{r}$ is mapped to a high-dimensional vector through the positional encoding, as follows:

$$\gamma(t, L) = (\sin(2^0 t\pi), \cos(2^0 t\pi), \ldots, \sin(2^L t\pi), \cos(2^L t\pi)) \tag{1}$$

where the scalar value t is an element of $\mathbf{x}$ and $\mathbf{r}$, and L is the number of frequency octaves.

Unlike Neural Radiance Fields (NeRFs) [38] that outputs an RGB color value, Neural Feature Fields have the potential to be utilized in various downstream tasks because it outputs a feature vector for a given 3D point and a ray direction. For the generative perspective, Niemeyer *et al.* [40] proposed a novel generative model for Neural Feature Fields. In their model, called GIRAFFE, the object representations are represented by Neural Feature Fields (denoted as h) parameterized by Multi-Layer Perceptron (MLP). In order to express different objects (in our case, *people with different poses and shapes*), the MLP is conditioned on latent vectors representing the object's shape ($\mathbf{z}_s$) and appearance ($\mathbf{z}_a$) as follows:

$$h : \mathbb{R}^{L_\mathbf{x}} \times \mathbb{R}^{L_\mathbf{r}} \times \mathbb{R}^{M_s} \times \mathbb{R}^{M_a} \to \mathbb{R}^+ \times \mathbb{R}^{M_f}$$
$$(\gamma(\mathbf{x}), \gamma(\mathbf{r}), \mathbf{z}_s, \mathbf{z}_a) \mapsto (\sigma, \mathbf{f}) \tag{2}$$

where $L_{\mathbf{x}}$ and $L_{\mathbf{r}}$ denote output dimensions of the positional encodings; M_s and M_a denote dimension of $\mathbf{z}_s$ and $\mathbf{z}_a$ respectively; σ and $\mathbf{f}$ denote a volume density and a feature vector. Finally, the model generates a realistic and controllable image through volume and neural rendering from inferred feature fields representing a specific scene.

In this work, we incorporate such a representation into conventional regression-based human mesh recovery so that the algorithm can implicitly imagine a person placed in a three-dimensional space.

3.2. SMPL Body Model

SMPL [34] is a parametric human body model. It provides a function $\mathcal{M}(\boldsymbol{\theta}, \boldsymbol{\beta})$ that takes pose and shape parameters (denoted as $\boldsymbol{\theta} \in \mathbb{R}^{72}$ and $\boldsymbol{\beta} \in \mathbb{R}^{10}$ respectively) as inputs and outputs a body mesh $M \in \mathbb{R}^{6890\times 3}$. The pose parameters consist of a global body rotation and 23 relative joint rotations. The shape parameters are the first 10 coefficients in the PCA shape space. For a given mesh, 3D joints J can be obtained by linear combination of the mesh vertices, $J = WM$, with pre-trained linear regressor W.

3.3. Model Architecture

The intuition behind our method is that if the model is trained to infer a human's pose and shape at arbitrary camera viewing directions in 3D space, the model learns spatial prior knowledge about human's appearance; consequently, the performance in the *canonical viewing direction* in which the image was taken is improved. To achieve this, as depicted in Fig. 2, our model mostly follows the HMR paradigm [19] where the input is an image of a person and output is the set of SMPL body model parameters, but there is a major difference in the *Feature Fields Module*. In this section, we first describe the operation of the Feature Fields Module consisting of *Feature Fields Generation* and *Volume Rendering*, and then explain *Parameter Regression* and *Geometric Guidance Branch* using the module.

Feature Fields Generation. The goal of the Feature Fields Module is to construct the person in 3D space at the feature-level so that the model can look at the person from an arbitrary viewing direction. Given an input image I, we first encode the image using the CNN Encoder g (*i.e.*, ResNet-50 [12] *before* the global average pooling) and obtain the feature vector $\mathbf{z} = g(I) \in \mathbb{R}^{2048\times 7\times 7}$. The encoded feature vector $\mathbf{z}$ may contain both information about the foreground and background of the given image. Thus, we use the *Foreground Attention* [56] $\mathcal{A}$ and obtain the human-related feature vector $\mathbf{z}_{\text{fg}} = GAP(\mathcal{A}(\mathbf{z})) \in \mathbb{R}^{2048}$, where $GAP(\cdot)$ denotes global average pooling. Finally, the MLP (denoted as h) representing feature fields is conditioned on the latent vector $\mathbf{z}_{\text{fg}}$, and implicitly expresses the human in a 3D space.

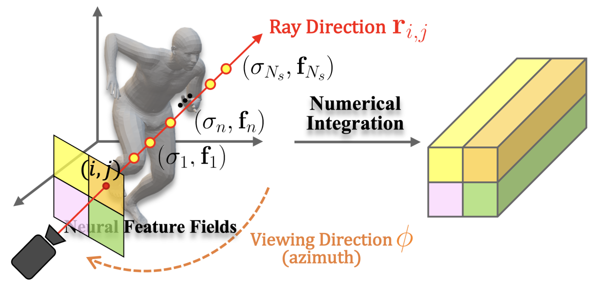

Figure 3. **Volume rendering procedure in a neural feature field.** To extract a 2D feature map from the Feature Field, sample points on the ray direction $\mathbf{r}_{i,j}$. From the volume density σ and feature vector $\mathbf{f}$, the 2D feature map is obtained by Numerical Integration.

Volume Rendering. In the feature field, we can look at the person represented by the feature field from an arbitrary viewing direction by controlling the camera pose (*i.e.*, camera extrinsic). For the camera pose, since the ambiguity of the human pose occurs in the horizontal direction, we fix the elevation of the camera pose to $0°$ and control only the azimuth (denoted as ϕ). For simplicity, we denote the camera pose as ϕ. Also, we define the direction in which the image was taken as *Canonical Viewing Direction* ($\phi = 0$).

To infer the human pose and shape viewed from a viewing direction ϕ, the 2D feature map $\mathbf{f}_\phi \in \mathbb{R}^{2048\times H\times W}$ should be obtained from the feature field by volume rendering [38], where H and W denote the spatial resolutions of the feature. Given the camera pose ϕ, let $\{\mathbf{x}_{i,j,n}\}_{n=1}^{N_s}$ be sample points on the ray direction $\mathbf{r}_{i,j}$ for the (i, j) location of the 2D feature map, where N_s is the number of sample points. We omit i and j for simplicity. Then, as shown in Fig. 3, we can obtain a feature vector $\mathbf{f}_n \in \mathbb{R}^{2048}$ and a volume density σ_n for each 3D point $\mathbf{x}$ as follows:

$$(\sigma_n, \mathbf{f}_n) = h(\gamma(\mathbf{x}_n), \gamma(\mathbf{r}), \mathbf{z}_{\text{fg}}). \tag{3}$$

where γ denotes the positional encoding.

Finally, using *Numerical Integration* as in [40], volume rendered feature vector $\mathbf{f}_{\text{rend}} \in \mathbb{R}^{2048}$ is obtained as follows:

$$\mathbf{f}_{\text{rend}} = \sum_{n=1}^{N_s} \tau_n \alpha_n \mathbf{f}_n \quad \tau_n = \prod_{k=1}^{n-1} (1 - \alpha_k) \quad \alpha_n = 1 - e^{-\sigma_n \delta_n} \tag{4}$$

where τ_n is the transmittance, α_n the alpha value for $\mathbf{x}_n$, and $\delta_n = \left|\left|\mathbf{x}_{n+1} - \mathbf{x}_n\right|\right|_2$ the distance between neighboring sample points. We repeat this process for each spatial location (i, j) and obtain the 2D feature map $\mathbf{f}_\phi \in \mathbb{R}^{2048\times H\times W}$, as in [40]. Finally, to use the feature map for SMPL parameter regression, we generate feature vector $\mathbf{z}_\phi = AGG(\mathbf{f}_\phi) \in \mathbb{R}^{2048}$, where $AGG(\cdot)$ denotes aggregation layer consisting of single depthwise convolution.

Parameter Regression. We have obtained the feature vector $\mathbf{z}_\phi$ that contains the information about the person

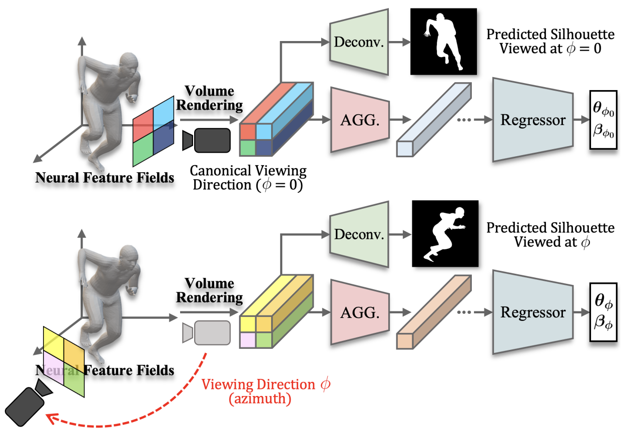

Figure 4. **SMPL parameter and silhouette regression with controlling camera viewing direction. Top:** regression from the *Canonical Viewing Direction* ($\phi = 0$), as in conventional methods. **Bottom:** regression from an arbitrary viewing direction.

viewed from the camera pose ϕ through *Feature Fields Module*. From the feature vector $\mathbf{z}_\phi$, ImpHMR predicts the SMPL model $\Theta_\phi = \{\boldsymbol{\theta}_\phi, \boldsymbol{\beta}_\phi, \boldsymbol{\pi}_\phi\}$ using the regressor $\mathcal{R}(\cdot)$ as $\Theta_\phi = \mathcal{R}(\mathbf{z}_\phi)$, where each element of Θ_ϕ denotes the pose, shape, and camera parameters inferred from the viewing direction ϕ respectively. Note that, when $\phi = 0$ (in short, ϕ_0), ImpHMR outputs Θ_{ϕ_0} that is the inference result viewed from the direction in which the image was taken, as in conventional regression-based methods, and otherwise outputs the inference result viewed from a viewing direction ϕ, as shown in Fig. 4. Therefore, after the training, we use the ϕ fixed to 0 for testing.

From the predicted parameters Θ, we can generate the body mesh with vertices $M = \mathcal{M}(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \mathbb{R}^{6890\times3}$. Subsequently, using the pre-trained linear regressor, 3D joints $J \in \mathbb{R}^{N_j\times3}$ ($\mathcal{J}_{3D}$ in Fig. 2) can be regressed from the mesh vertices M, where N_j is the number of joints. Furthermore, 2D keypoints $K \in \mathbb{R}^{N_j\times2}$ ($\mathcal{J}_{2D}$ in Fig. 2) are obtained as $K = \boldsymbol{\Pi}(J)$, where $\boldsymbol{\Pi}(\cdot)$ denotes the projection function from weak-perspective camera parameters $\boldsymbol{\pi} \in [s, t]$ (s and t denote the scale and translation parameters, respectively).

Geometric Guidance Branch. ImpHMR is trained to regress the rotated ground-truth viewed from an arbitrary viewing direction ϕ to learn spatial prior of human's appearance (see Sec. 3.4). However, unlike GIRAFFE, which generates images, our model regresses parameters (*i.e.*, SMPL), so there might not be enough information for the model to learn the geometry of 3D space. Thus, as shown in Fig. 4, we have the model reconstruct the silhouette S_ϕ (viewed at direction ϕ) from the 2D feature map $\mathbf{f}_\phi$ using deconvolution $\mathcal{D}(\cdot)$ as $S_\phi = \mathcal{D}(\mathbf{f}_\phi)$. To explicitly give geometric supervision for *unseen-view*, we generate the G.T. silhouette from the G.T. SMPL mesh rotated by the viewing direction using NMR [21], as shown in Fig. 5. Note that, the geometric guidance branch is used *only* for training.

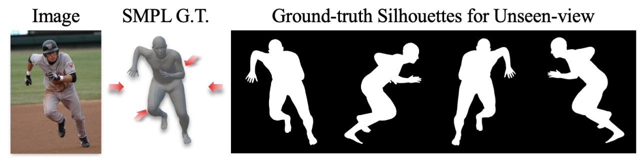

Figure 5. **Generating ground-truth silhouettes viewed from unseen-view.** G.T. silhouette from an arbitrary viewing direction is generated by rotating the mesh of SMPL G.T. and rendering it.

3.4. Training Objective

The final goal of our method is to improve the regression performance in the *canonical viewing direction* (ϕ_0) by having the model learn spatial prior about the person in 3D space. In this section, we describe the following three objectives for training: *Canonical View Regression*, *Arbitrary View Imagination*, and *Appearance Consistency Loss*.

Canonical View Regression Loss. This is the constraint for inference from *canonical viewing direction* (ϕ_0) just like previous methods [19, 25]. 2D keypoints K_{ϕ_0} and 3D joints J_{ϕ_0} are obtained from inferred SMPL parameters (*i.e.*, $\boldsymbol{\theta}_{\phi_0}$, $\boldsymbol{\beta}_{\phi_0}$, and $\boldsymbol{\pi}_{\phi_0}$), making them close to their G.T. as follows:

$$\begin{aligned}\mathcal{L}_{reg} = \lambda_{2d}||K_{\phi_0} - \hat{K}|| + \lambda_{3d}||J_{\phi_0} - \hat{J}|| \\ + \lambda_{pose}||\boldsymbol{\theta}_{\phi_0} - \hat{\boldsymbol{\theta}}|| + \lambda_{shape}||\boldsymbol{\beta}_{\phi_0} - \hat{\boldsymbol{\beta}}||,\end{aligned} \tag{5}$$

where $||\cdot||$ is the squared L2 norm; $\hat{K}$, $\hat{J}$, $\hat{\boldsymbol{\theta}}$, and $\hat{\boldsymbol{\beta}}$ denote the ground-truth 2D keypoints, 3D joints, and SMPL pose and shapes, respectively following the notation of [62].

Arbitrary View Imagination Loss. To leverage the consistency of pose and shape from unseen-views, we train the model to infer human's appearance viewed at arbitrary directions. Thus, if *3D labels exist*, we use the constraint that the predicted result from an arbitrary viewing direction ϕ (sampled from the distribution of camera pose p_{cam}) should be equal to the ground-truth rotated by $-\phi$ as follows:

$$\begin{aligned}\mathcal{L}_{imag} = \mathbb{E}_{\phi\sim p_{cam}}[\lambda_{3d}||J_\phi - \hat{J}_{-\phi}|| + \lambda_{silh.}||S_\phi - \hat{S}_{-\phi}|| \\ + \lambda_{pose}||\boldsymbol{\theta}_\phi - \hat{\boldsymbol{\theta}}_{-\phi}|| + \lambda_{shape}||\boldsymbol{\beta}_\phi - \hat{\boldsymbol{\beta}}||],\end{aligned} \tag{6}$$

where $\hat{J}_{-\phi}$ is ground-truth 3D joints rotated by $-\phi$ in horizontal direction, $\hat{S}_{-\phi}$ is G.T. silhouette viewed at $-\phi$, and $\hat{\boldsymbol{\theta}}_{-\phi}$ is ground-truth pose rotated by $-\phi$ only for the global orientation, and $p_{cam} \sim \mathcal{U}[0, 2\pi]$. Using the constraint, we can disentangle human appearance and viewing direction, resulting in better spatial prior about humans in 3D space.

Note that we rotate the ground-truth by $-\phi$ because the viewing direction and shown person's appearance are rotated oppositely. In addition, for the ground-truth shape parameters $\hat{\boldsymbol{\beta}}$, we do not apply the rotation because it is independent of the viewing direction.

Appearance Consistency Loss. ImpHMR can make predictions in various viewing directions from a single image. By utilizing this capability of the model, we perform *self-supervised* learning through the constraint that SMPL parameters inferred from different directions (sampled from p_{cam}) should be the same if there are *only 2D labels* as:

$$\mathcal{L}_{cons} = \mathbb{E}_{\phi_1,\phi_2 \sim p_{cam}}[\lambda_{pose}||\boldsymbol{\theta}'_{\phi_1} - \boldsymbol{\theta}_{\phi_2}|| + \lambda_{shape}||\boldsymbol{\beta}_{\phi_1} - \boldsymbol{\beta}_{\phi_2}||], \quad (7)$$

where $\boldsymbol{\theta}'_{\phi_1}$ denotes the modified parameters where only the global orientation of the inferred pose parameters $\boldsymbol{\theta}_{\phi_1}$ is changed by $\phi_2 - \phi_1$ amount. Finally, our overall loss function is $\mathcal{L}_{all} = \mathcal{L}_{reg} + \mathcal{L}_{imag} + \mathcal{L}_{cons}$. We selectively use each loss function depending on whether 3D labels are available or not and our model is trained *end-to-end* manner.

4. Experiments

4.1. Datasets and Evaluation Metrics

Following previous works [19, 24, 25], we use a mixture of 2D and 3D datasets. We use MPI-INF-3DHP [36] and Human3.6M [13] with ground-truth SMPL as our 3D datasets for training. Also, MPII [1], COCO [31], and LSPET [16] with the pseudo-ground-truth SMPL provided by [17] are used as 2D datasets. As in PARE [24], we divide the training process into two phases to reduce the overall training time. We first train our model on COCO for ablation studies, and then obtain the final performance using a mixture of all datasets for comparison with SOTA methods. For evaluation, we use 3DPW [54] and 3DPW-OCC [63] for quantitative evaluation. Our method is evaluated using mean per joint position error (MPJPE), Procrustes-aligned mean per joint position error (PA-MPJPE), and per-vertex error (PVE) metrics. For qualitative evaluation, we evaluate the quality of the inferred mesh on 3DPW, LSP [15], and COCO validation sets. More description about datasets is in the supplementary material.

4.2. Experimental Results

In this section, we validate the effectiveness of the proposed method. First, we compare the performance of ImpHMR with previous SOTA methods. Then, we confirm whether ImpHMR has the ability to infer human appearance viewed from different viewing directions in 3D space. Finally, the efficacy of each of the methods is validated.

Comparison with State-of-the-Art. First, we evaluate the human mesh recovery performance of ImpHMR. Table 1 shows the quantitative results of previous state-of-the-art and our method on 3DPW test split. As shown in Tab. 1, our method (denoted as "ImpHMR (Ours)") shows superior performance compared to other methods for all metrics in both temporal- and frame-based approaches. In particular,

		3PDW		
	Method	MPJPE ↓	PA-MPJPE ↓	PVE ↓
Temporal	HMMR [20]	116.5	72.6	139.3
	DSD [50]	-	69.5	-
	Arnab *et al.* [2]	-	72.2	-
	Doersch *et al.* [11]	-	74.7	-
	VIBE [23]	93.5	56.5	113.4
	TCMR [8]	95.0	55.8	111.3
	MPS-Net [55]	91.6	54.0	109.6
Frame-based	HMR [19]	130.0	76.7	-
	GraphCMR [26]	-	70.2	-
	SPIN [25]	96.9	59.2	116.4
	PyMAF [62]	92.8	58.9	110.1
	I2L-MeshNet [39]	100.0	60.0	-
	ROMP [49]	89.3	53.5	105.6
	HMR-EFT [17]	-	54.2	-
	PARE [24]	<u>82.9</u>	<u>52.3</u>	<u>99.7</u>
	ImpHMR (Ours)	**81.8**	**49.8**	**96.4**
	ImpHMR (Ours) w. 3DPW	74.3	45.4	87.1

Table 1. **Results on 3DPW.** Best in bold, second-best underlined. Values are in mm. "ImpHMR (Ours)" and "ImpHMR (Ours) w. 3DPW" denote the model trained *w/o* and *w.* 3DPW train set, respectively.

Figure 6. **Qualitative results.** Qualitative comparison of the proposed method with SPIN [25] and HMR-EFT [17] on COCO validation set and 3DPW test split.

ImpHMR shows an -4.4mm (8.1%) performance improvement in PA-MPJPE metric compared to HMR-EFT [17]. Also, it shows a -1.1mm (1.3%), -2.5mm (4.8%), and -3.3mm (3.3%) improvement in MPJPE, PA-MPJPE, and PVE, respectively, compared to PARE [24], the most previous best performing model. Also, we report the results of the model trained using 3DPW train split (denoted as "ImpHMR (Ours) w. 3DPW" in Tab. 1) to see the performance of the model when using the ground-truth SMPL labels. Compared to when the dataset is not used, ImpHMR shows a significant performance improvement and outperforms all methods by a large margin.

We perform an evaluation on 3DPW-OCC [63], an

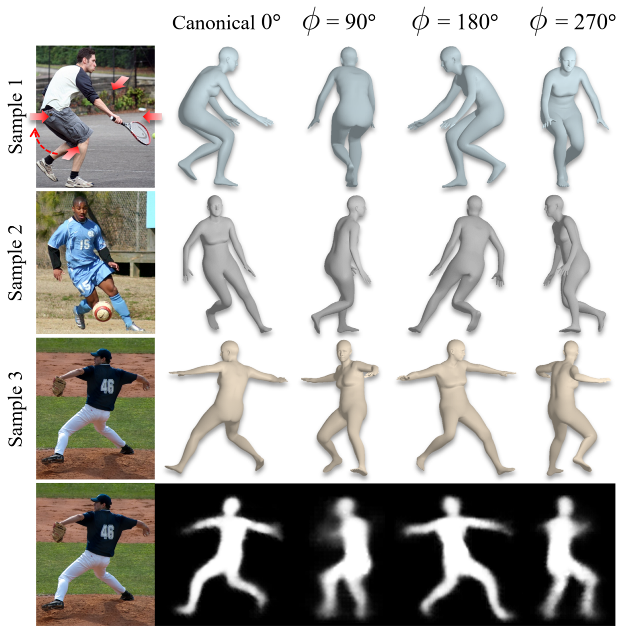

Figure 7. **Inferred SMPL mesh and silhouettes viewed from different viewing directions.** Results inferred *by changing the viewing direction* clockwise by 90° from canonical viewing direction. Note that the inference results are *not by rotating the mesh* inferred from the canonical viewing direction, but *directly inferring a person viewed from different directions in 3D space*.

occlusion-specific dataset, to verify the performance of ImpHMR in the presence of ambiguity (*e.g.*, occlusion). Table 2 shows the result. For fair comparison, all methods in the table are trained on the same datasets (*i.e.*, Human3.6M [13], COCO [31], and, 3DOH [63]). As shown in Tab. 2, ImpHMR outperforms the occlusion-specific methods Zhang *et al*. [63] and PARE [24], including HMR-EFT [17], by a large margin. This demonstrates that the structure and learning method of ImpHMR is suitable for modeling situations in which ambiguity is present.

For qualitative comparisons, we compare our method with SPIN [25] and HMR-EFT [17]. As shown in the Fig. 6, ImpHMR outputs a mesh that is well aligned with the image even when a person with extreme poses or ambiguity exists.

Results from Different Viewing Directions. In this section, we verify that ImpHMR successfully imagines a person in 3D space. To do this, we report the mesh reconstruction results from different viewing directions ϕ (*i.e.*, 0°, 90°, 180°, and 270°) for a given image of a person. As shown in Fig. 7, we confirm that ImpHMR can imagine a person's appearance not only from the canonical viewing direction, but also from the image from the left, right, and back of the person. Note that the inference results are *not by rotating the mesh* inferred from the canonical viewing direction, but the results of *directly inferring a person viewed from different directions in 3D space*. A viewing direction can be an arbitrary angle; herein, we report only the results

Method	MPJPE ↓	PA-MPJPE ↓	PVE ↓
Zhang *et al*. [63]	-	72.2	-
HMR-EFT [17]	94.4	60.9	111.3
PARE [24]	90.5	56.6	107.9
ImpHMR (Ours)	**86.5**	**54.4**	**104.7**

Table 2. **Results on 3DPW-OCC.** For fair comparison, all methods are trained using the same datasets (*i.e.*, Human3.6M, COCO, and 3DOH). Best in bold.

Method	SPL. 1	SPL. 2	SPL. 3	LSP dataset
$\mathcal{L}_{reg}$	0.0763	0.0875	0.0711	0.0844
$\mathcal{L}_{reg} + \mathcal{L}_{imag}$	0.0618	0.0512	0.0639	0.0774
$\mathcal{L}_{reg} + \mathcal{L}_{imag} + \mathcal{L}_{cons}$	0.0361	0.0314	0.0445	0.0561

Table 3. **Entanglement between shape and viewing direction.** Each value denotes the degree of *variation of the inferred shape* when inferred by changing the viewing direction. All methods are trained using COCO dataset.

	Method	MPJPE ↓	PA-MPJPE ↓	PVE ↓
	Baseline	101.6	58.3	117.2
w/o $\mathcal{D}(\cdot)$	$\mathcal{L}_{reg}$	96.5	58.9	116.0
w/o $\mathcal{D}(\cdot)$	$\mathcal{L}_{reg} + \mathcal{L}_{imag}$	94.9	57.9	114.2
w/o $\mathcal{D}(\cdot)$	$\mathcal{L}_{reg} + \mathcal{L}_{imag} + \mathcal{L}_{cons}$	93.5	57.5	113.1
	$\mathcal{L}_{reg} + \mathcal{L}_{imag} + \mathcal{L}_{cons}$	92.7	57.0	112.1

Table 4. **Effectiveness of each proposed method.** The results are evaluated on the 3DPW dataset. Values for all metrics are in mm. w/o $\mathcal{D}(\cdot)$ denotes the method trained without Geometric Guidance Branch. All methods are trained using COCO dataset.

from the 4 different angles.

Additionally, the last row of Fig. 7 is the person's silhouettes in Sample 3 viewed from various directions by using $\mathcal{D}(\cdot)$, and it can be seen that the inferred silhouettes are similar to what a human imagines. This demonstrates that spatially meaningful information is contained in a volume-rendered 2D feature map $\mathbf{f}_\phi$ by an appropriate guide of Geometric Guidance Branch.

To quantitatively verify the 3D spatial construction capability of ImpHMR, we measure the **E**ntanglement between the **S**hape and the **V**iewing direction (in short, ESV). If the 3D space is well constructed in Neural Feature Fields, the body shape should be consistent despite changes in viewing direction. Therefore, we change the viewing direction from 0° to 360° at 1° intervals and define the average of the standard deviations of the inferred shape parameters as ESV, which is the degree of entanglement. We measure the ESV for each sample image in Fig. 7 and the *entire* LSP [15] dataset. As shown in Tab. 3, we can notice that the degree of entanglement decreases as the proposed constraints are added. This indicates that ImpHMR successfully disentangles body shape and viewing direction, as a result imagining a person in a 3D space well. A detailed description of ESV is in the supplementary material.

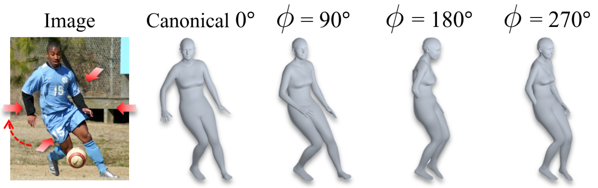

Figure 8. **Inferred SMPL mesh from different viewing directions with *Explicit* representation.** Results inferred by changing the direction clockwise by 90° from canonical viewing direction.

Ablation Studies. To verify the efficacy of each of the proposed methods, we evaluate the performance change by adding each method. For a fair comparison, we train a baseline model (denoted as "Baseline" in Tab. 4) that has the same model architecture and the number of parameters as ImpHMR (except $AGG(\cdot)$) but does not perform feature fields generation and volume rendering. As shown in Tab. 4, we can notice that all methods provide a positive contribution. Compared to Baseline, it can be seen that there is an improvement even when just generating a feature vector through volume rendering within feature fields (denoted as $\mathcal{L}_{reg}$). This verifies the inference method of ImpHMR is suitable for HMR tasks. In addition, as shown in Tab. 3 and Tab. 4, by adding the proposed constraints, including silhouette loss with *geometric guidance branch*, the better the model disentangles the person's appearance and viewing direction in 3D space, and the performance increases accordingly. Through this, we can confirm that our assumption about the proposed method is valid.

Table 5 shows the ablation of the model architecture (*i.e.*, aggregation layer AGG and foreground attention $\mathcal{A}$). We use three types of AGG: global average pooling (GAP), convolution (Conv.), and depth-wise convolution (DWConv.), and report the performance of combinations with $\mathcal{A}$. As can be seen in Tab. 5, AGG shows good performance in the order of DWConv, GAP, and Conv. We can notice that foreground attention has a positive effect except for Conv. We finally adopted the best-performing set of Tab. 5 (e).

ImpHMR uses neural feature fields, an implicit representation, to imagine a person in 3D. However, as a means of expressing 3D space, there is also an explicit representation such as the voxel-based method (*e.g.*, PTN [59]). To explore the suitability of explicit representation, we check the performance of the baseline in which Feature Fields Module of ImpHMR is replaced by a voxel-based representation. For the volumetric representation, we use Perspective Transformer Nets [59] (PTN). For a fair comparison, we set the voxel resolution to $4\times4\times4$, the same as ImpHMR. Since PTN can perspective project features for a given camera extrinsic, we train the baseline using the same constraints (*i.e.*, $\mathcal{L}_{reg}$, $\mathcal{L}_{imag}$, and $\mathcal{L}_{cons}$) as in ImpHMR. Figure 8 shows the inference results of the baseline, and we can notice that it fails to model 3D space. This indicates that implicit representation is more suitable for modeling a person in 3D.

Method		MPJPE ↓	PA-MPJPE ↓	PVE ↓
	HMR-EFT [17]	99.0	59.9	-
AGG **Aggregation**	$\mathcal{A}$ **Attention**			
(a) GAP	with $\mathcal{A}$	95.3	57.6	114.5
(b) GAP	without $\mathcal{A}$	95.8	57.9	114.9
(c) Conv.	with $\mathcal{A}$	96.0	58.0	114.6
(d) Conv.	without $\mathcal{A}$	95.2	57.7	114.9
(e) DWConv.	with $\mathcal{A}$	**92.7**	**57.0**	**112.1**
(f) DWConv.	without $\mathcal{A}$	94.7	57.9	113.8

Table 5. **Ablation study of the model architecture.** "AGG Aggregation" denotes the type of aggregation layer in volume rendering. "$\mathcal{A}$ Attention" denotes whether foreground attention is used. All methods are trained using COCO dataset.

Method	Res.1	Res.2	Res.4	Res.6
HMR-EFT [17]	115.5	-	-	-
PyMAF [62]	33.6	-	-	-
PARE [24]	27.5	-	-	-
ImpHMR (Ours)	88.8	88.4	87.1	78.2

Table 6. **Comparison of inference speed.** The numbers are in *frames per second* (**fps**). The Res. denotes the spatial resolution of a 2D feature map in volume rendering for our method. Thanks to efficient spatial representation in feature fields, ImpHMR shows about $2 \sim 3$ *times faster* fps compared to PyMAF and PARE.

Table 6 compares the inference speed between ImpHMR and the current SOTA methods. For fair evaluation, frames per second (fps) is calculated by averaging the time it took for each model to infer 10000 times of an input image of 224×224 size on RTX 2080Ti GPU. As shown in Tab. 6, we can notice that ImpHMR is slightly slower than HMR-EFT, but still has *real-time* performance. Especially, ImpHMR has $2 \sim 3$ *times faster* fps than PyMAF and PARE, which are current SOTA methods. This is because ImpHMR is capable of efficient spatial representation within neural feature fields compared to the latest SOTA methods that utilize spatial information.

5. Conclusion and Future Works

We have introduced a novel HMR model called "ImpHMR" that can implicitly imagine a human in 3D space from a given 2D observation via neural feature fields. To utilize consistency with pose and shape from unseen-views, we propose arbitrary view imagination loss and appearance consistency loss. Also, we propose geometric guidance branch that helps the model can learn better geometric information. ImpHMR has $2 \sim 3$ times faster fps than current SOTAs thanks to efficient spatial representation in feature fields. Also, extensive evaluation proves that our method is valid. For future works, we can make a more occlusion-robust model by carefully modeling volume density.

Acknowledgements. This work was conducted by Center for Applied Research in Artificial Intelligence (CARAI) grant funded by DAPA and ADD (UD190031RD).

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 6

[2] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 6

[3] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter V. Gehler, Javier Romero, and Michael J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016. 2

[4] Wenzheng Chen, Huan Ling, Jun Gao, Edward Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to predict 3d objects with an interpolation-based differentiable renderer. *Advances in Neural Information Processing Systems*, 32:9609–9619, 2019. 3

[5] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. 3

[6] Hanbyel Cho, Jaesung Ahn, Yooshin Cho, and Junmo Kim. Video inference for human mesh recovery with vision transformer. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–6, 2023. 1, 2

[7] Hanbyel Cho, Yooshin Cho, Jaemyung Yu, and Junmo Kim. Camera distortion-aware 3d human pose estimation in video with optimization-based meta-learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11169–11178, 2021. 1

[8] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1964–1973, June 2021. 1, 2, 6

[9] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 769–787, Cham, 2020. Springer International Publishing. 2

[10] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 20–40, Cham, 2020. Springer International Publishing. 2

[11] Carl Doersch and Andrew Zisserman. Sim2real transfer learning for 3d human pose estimation: motion to the rescue. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 6

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 4

[13] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 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, jul 2014. 6, 7

[14] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[15] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proc. BMVC*, pages 12.1–11, 2010. doi:10.5244/C.24.12. 6, 7

[16] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR 2011*, pages 1465–1472, 2011. 6

[17] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. In *3DV*, 2020. 2, 6, 7, 8

[18] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8320–8329, 2018. 2

[19] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Regognition (CVPR)*, 2018. 1, 2, 4, 5, 6

[20] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2, 6

[21] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5

[22] Rawal Khirodkar, Shashank Tripathi, and Kris Kitani. Occluded human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1715–1725, June 2022. 1, 2

[23] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 6

[24] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Proc. International Conference on Computer Vision (ICCV)*, pages 11127–11137, Oct. 2021. 2, 6, 7, 8

[25] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 1, 2, 5, 6, 7

[26] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 6

[27] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *ICCV*, 2021. 1

[28] Jogendra Nath Kundu, Mugalodi Rakesh, Varun Jampani, Rahul M Venkatesh, and R. Venkatesh Babu. Appearance consensus driven self-supervised human mesh recovery. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[29] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6050–6059, 2017. 2

[30] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. 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*, pages 3383–3393, 2021. 1, 2

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and Larry Zitnick. Microsoft coco: Common objects in context. In *ECCV*. European Conference on Computer Vision, September 2014. 6, 7

[32] Ruixu Liu, Ju Shen, He Wang, Chen Chen, Sen-ching Cheung, and Vijayan Asari. Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5064–5073, 2020. 1

[33] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7708–7717, 2019. 3

[34] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 1, 2, 3, 4

[35] Zhengyi Luo, S. Alireza Golestaneh, and Kris M. Kitani. 3d human motion estimation via motion compression and refinement. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020. 1, 2

[36] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017. 6

[37] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 3

[38] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020. 3, 4

[39] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 752–768. Springer, 2020. 2, 6

[40] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11453–11464, June 2021. 1, 3, 4

[41] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V. Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. Verona, Italy, 2018. 2

[42] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[43] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. Texturepose: Supervising human mesh estimation with texture consistency. In *ICCV*, 2019. 2

[44] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[45] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. 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 (CVPR)*, June 2018. 2

[46] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7753–7762, 2019. 1

[47] Yu Rong, Ziwei Liu, Cheng Li, Kaidi Cao, and Chen Change Loy. Delving deep into hybrid annotations for 3d human recovery in the wild. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2

[48] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3

[49] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, 2021. 1, 2, 6

[50] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 6

[51] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *NeurIPS*, 2020. 3

[52] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5236–5246. Curran Associates, Inc., 2017. 2

[53] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 20–36, 2018. 2

[54] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 6

[55] Wen-Li Wei, Jen-Chun Lin, Tyng-Luh Liu, and Hong-Yuan Mark Liao. Capturing humans in motion: Temporal-attentive 3d human pose and shape estimation from monocular video. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 1, 2, 6

[56] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 4

[57] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 1

[58] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7760–7770, 2019. 2

[59] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 2, 8

[60] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, Bill Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In *European Conference on Computer Vision (ECCV)*, pages 465–481, 2020. 2

[61] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2148–2157, 2018. 2

[62] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 1, 2, 5, 6, 8

[63] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 6, 7

[64] Yuxiang Zhang, Zhe Li, Liang An, Mengcheng Li, Tao Yu, and Yebin Liu. Lightweight multi-person total motion capture using sparse multi-view cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5560–5569, 2021. 2