Submitted to ICCV 2023

3D-Aware Neural Body Fitting for Occlusion Robust 3D Human Pose Estimation

Yi Zhang[1*] Pengliang Ji[2*] Adam Kortylewski[3,4] Angtian Wang[1]
Jieru Mei[1] Alan Yuille[1]
[1]Johns Hopkins University [2]Beihang University
[3]Max Planck Institute for Informatics [4]University of Freiburg

Abstract

Regression-based methods for 3D human pose estimation directly predict the 3D pose parameters from a 2D image using deep networks. While achieving state-of-the-art performance on standard benchmarks, their performance degrades under occlusion. In contrast, optimization-based methods fit a parametric body model to 2D features in an iterative manner. The localized reconstruction loss can potentially make them robust to occlusion, but they suffer from the 2D-3D ambiguity. Motivated by the recent success of generative models in rigid object pose estimation, we propose 3D-aware Neural Body Fitting (3DNBF) - an approximate analysis-by-synthesis approach to 3D human pose estimation with SOTA performance and occlusion robustness. In particular, we propose a generative model of deep features based on a volumetric human representation with Gaussian ellipsoidal kernels emitting 3D pose-dependent feature vectors. The neural features are trained with contrastive learning to become 3D-aware and hence to overcome the 2D-3D ambiguity. Experiments show that 3DNBF outperforms other approaches on both occluded and standard benchmarks.

1. Introduction

Monocular 3D human pose estimation (HPE) is a long-standing problem of computer vision. Regression-based methods [11, 23, 48, 52, 65, 67] directly regress the 3D pose parameters of a human body model, such as SMPL [39], and learn to overcome the inherent 2D-3D ambiguity of the prediction task from the training data. However, the performance of regression-based methods degrades when humans are partially occluded, as demonstrated by related work [25] and in our experiments (Figure 1 (c)). Optimization-based methods [4, 26, 50, 73, 75] fit a parametric body model to 2D representations, such as keypoint detections [23, 27] or segmentation maps [48, 52, 77], in an iterative manner. They are relatively robust to occlusion but perform worse than regression-based methods in 3D HPE particularly because they suffer from the 2D-3D ambiguity (Figure 1(d)), even when regularized with strong 3D priors [50], because the manually designed 2D features lack 3D information.

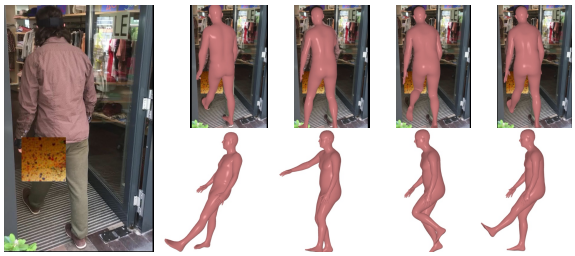

Figure 1: 3D human pose estimation under occlusion. Performance of regression-based methods [21] degrades under occlusion (c). Traditional optimization-based methods can be robust occlusion, but they suffer from the 2D-3D ambiguity in monocular 3D HPE (d). Our generative approach resolves the 2D-3D ambiguity through analysis-by-synthesis in a 3D-aware feature space (e).

Recently, generative models have been shown to be successful with improved robustness to occlusion in object recognition [30] and rigid object pose estimation [69] for certain object categories. The idea is to formulate vision tasks as inverse graphics or analysis-by-synthesis [24, 76] - searching for the parameters in a generative model (e.g. computer graphics models) that best explain the observed image while an outlier process can be introduced to explain occluded regions. However, performing analysis-by-synthesis in RGB pixel space is challenging due to the lack of both good generative models and efficient algorithms to inverse them. Instead, they perform approximate analysis-by-synthesis in deep feature space. However, the generative models used are 2D-based or simple cuboid-like 3D structures with features invariant to the 3D viewpoint, making them less suitable for 3D HPE.

*Indicates equal contribution.

In this work, we propose a 3D-aware Neural Body Fitting (3DNBF) framework that enables feature-level analysis-by-synthesis for 3D HPE, which is highly robust to occlusion (Figure 1(e)). Specifically, we propose a novel generative model of deep network features for human body, named Neural Body Volumes (NBV). NBV is an explicit volume-based parametric body representation consisting of a set of Gaussian ellipsoidal kernels that emit feature vectors. Compared with the popular mesh representation, our volume representation is analytically differentiable, provides smooth gradients, i.e. is efficient to optimize, and rigorously handles self-occlusion [55]. We employ a factorized likelihood model for feature maps which is further made robust to partial occlusion by incorporating robust loss functions [17]. To overcome the 2D-3D ambiguity, we impose a distribution on the kernel features conditioned on pose parameters making them pose-dependent.

Unlike optimization-based methods that manually design the feature representation which may lose information, we learn the features from data. In particular, we introduce a contrastive learning framework [2, 12, 72] to learn features that are invariant to instance-specific details (such as color of the clothes), meanwhile encouraging them to capture local 3D pose information of the human body parts, i.e. being 3D-aware. The generative model is learned with the feature extractor network iteratively. For more efficient inference, we attach a regression head to the feature extractor to predict the pose and shape parameters from the feature directly. During inference, we initialize NBV with the prediction from our regressor head and optimize the human pose by maximizing the likelihood of the target feature map under the generative model using gradient-based optimization. We find this combined approach can resolve common errors of regression-based methods, such as when the pose of partially occluded parts is not estimated correctly (Figure 1).

We evaluate 3DNBF on three existing 3D HPE datasets: `3DPW` [68], `3DPW-Occ` [78] and `3DOH50K` [78], and propose a more challenging adversarial evaluation protocol `3DPW-AdvOcc` for occlusion robustness. Our experimental results show that 3DNBF outperforms state-of-the-art (SOTA) regression-based methods as well as optimization-based approaches by a large margin under occlusion while maintaining SOTA performance on non-occluded data. In summary, our main contributions are:

1. We propose 3DNBF - an approximate analysis-by-synthesis approach for 3D HPE at feature level with a volume-based neural generative model NBV for human body with pose-dependent kernel features.

2. We introduce a contrastive learning framework to train NBV with a feature extractor such that the feature activations capture the local 3D-pose information of the body parts, to resolve the 2D-3D ambiguity.

3. We demonstrate on four datasets that 3DNBF outperforms SOTA regression-based and optimization-based methods, particularly when under occlusion.

2. Related Work

Monocular 3D Human Pose Estimation. Existing approaches can be categorized into regression-based and optimization-based methods. Regression-based methods [11, 23, 48, 52, 65, 67] directly estimate 3D human pose from RGB image using a deep network. Different 3D human pose representations are adopted such as 3D joint locations [40, 57], 3D heatmaps [51, 64, 79] and parameters of a parametric human body [23, 27, 52]. To model pose ambiguities, e.g. for truncated human images, [3, 28] predict multiple possible poses that have correct 2D projections. Optimization-based methods [4, 26, 50, 73, 75] involve parametric human models like SMPL [1, 39, 50], and produce both the 3D human pose and human shape. The representative method is SMPLify [4], which fits the SMPL model to 2D keypoint detections with strong priors. Exploiting more information into the fitting procedure has been investigated, including silhouettes [32], multi-view [16], more expressive shape models [22]. [73] propose to fit 3D part affinity maps to overcome 2D-3D ambiguity. This requires the network to learn accurate part orientation which is difficult and shown to be less robust to occlusion.

Robustness to Occlusion. Regression-based methods are sensitive to occlusions as studied by Kocabas *et al*. [25], who propose part segmentation guided attention mechanism to handle occlusion. Data augmentation is another common way to enhance occlusion robustness, for example by cropping [3, 21, 56], or by putting patches into the image [10, 61]. Even with data augmentation, we show that they are still sensitive to occlusion by applying a more sophisticated sliding-window attack. Explicit occlusion handling in regression-based methods infers occluded joints using representation redundancy [41, 42] which is partially successful or exploits visibility information in the training [7, 37, 70, 78]. However, the occlusion information except self-occlusion is often unavailable in the wild and expensive to annotate which limits the applicability of such methods. Another direction to handle occlusion is to use motion in sequences [6, 15]. Generative models are shown to be robust to occlusion [43, 69] for parsing rigid objects and we further demonstrate it for articulated objects.

Human body representations. Parametric mesh models [4, 26, 50, 73, 75] are most popular models for human pose/shape estimation, which generate intermediate representations like 2D keypoints, sihouette and part segmentations. However, these representations lose the local information, e.g. shading, useful for inferring 3D from 2D. Recently, implicit volume representation has become increasingly popular [20, 36, 47, 49, 53, 60, 63, 71, 74] as they

can achieve highly realistic human reconstruction. However, they are not suitable for our purpose as training these models often requires multi-view or videos and takes an extended time for a single person. We propose a body representation that combines a volumetric 3D Gaussian representation [55] which gives more stable gradients compared to mesh-based differentiable rendering. Compared to popular implicit volume representations, our volume representation is explicit with fewer parameters to learn which leads to efficient inference.

Generative Models of Neural Textures. Prior works have shown the potential of combining 3D representations with neural texture maps, with an application to image synthesis of static scenes through neural rendering [45, 46, 66]. As inverting a generative model of RGB pixel values is challenging, a recent line of work introduced a neural analysis-by-synthesis approach to perform visual recognition tasks such as image classification [29–31] and 3D pose estimation [69] with a largely enhanced robustness to partial occlusion when compared to standard deep network based approaches. However, these prior works explicitly assume rigid objects and use simple 2D-based or cuboid-like generative models. [44] learns a continuous feature embedding function for each vertex on 3D human body mesh. However, this representation is invariant to 3D pose and therefore loses the useful information for estimation 3D from 2D. Our work generalizes the neural analysis-by-synthesis approach to 3D HPE, addressing the challenge of 2D-3D ambiguities and modeling articulated human bodies.

3. 3D-Aware Neural Body Fitting

In the following, we first explain a conceptual formulation of analysis-by-synthesis for 3D human pose estimation (Section 3.1) and propose our feature-level analysis-by-synthesis formulation in Section 3.2. Then we introduce the proposed generative Neural Body Volume model (Section 3.3), including the 3D-aware pose-dependent features (Section 3.4). Finally, we describe the training and inference process for the generative model in Section 3.5, 3.6.

3.1. HPE via Analysis-by-Synthesis

Given an input image $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 3}$, we aim to estimate the 3D human pose parameters $\boldsymbol{\theta}$. Using Bayes rule we formulate the pose estimation task as a probabilistic inference problem given the observed image $\boldsymbol{I}$:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\boldsymbol{\theta}|\boldsymbol{I}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\boldsymbol{I}|\boldsymbol{\theta})p(\boldsymbol{\theta}), \qquad (1)$$

where $p(\boldsymbol{\theta})$ is a prior distribution learned from data [4, 50], and $p(\boldsymbol{I}|\boldsymbol{\theta})$ is the likelihood. $p(\boldsymbol{I}|\boldsymbol{\theta})$ is typically defined using a generative forward model (involving 3D CAD models and a graphics engine) , and the analysis-by-synthesis process is hence defined as finding the parameters $\boldsymbol{\theta}^*$ that can best explain the input image. However, it is very challenging to reconstruct human images accurately which requires either multi-view images or video input [20, 36, 53, 63, 71].

Instead of performing analysis-by-synthesis in RGB space, we aim to reconstruct the human appearance at the *feature-level* of a neural network. Fig. 2 is an overview of our method. The feature representations will be learned to become invariant to image variations that is not relevant for the HPE task, such as clothing color or style, and hence will enable us to perform HPE accurately from a single image. In the following, we will first introduce the concept of feature-level analysis-by-synthesis and subsequently introduce a generative model of humans on the feature level.

3.2. Feature-Level Analysis-by-Synthesis

We denote a feature representation of an input image as $\zeta(\boldsymbol{I}) = \boldsymbol{F} \in \mathbb{R}^{H\times W\times D}$ which is the output of a deep convolutional neural network ζ. $\boldsymbol{f}_i \in \mathbb{R}^D$ is a feature vector in $\boldsymbol{F}$ at pixel i on the feature map. We define a generative model of humans on the feature-level as $\mathcal{G}(\boldsymbol{\theta}) = \hat{\boldsymbol{\Phi}} \in \mathbb{R}^{H\times W\times D}$, which produces a feature map $\hat{\boldsymbol{\Phi}}$ given the pose θ. We can now define the likelihood function of our Bayesian model (Eq. 1). To enable efficient learning and inference, we adopt a factorized likelihood model:

$$p(\boldsymbol{F}|\mathcal{G}(\boldsymbol{\theta}), \mathcal{B}) = \prod_{i\in\mathcal{FG}} p(\boldsymbol{f_i}|\hat{\boldsymbol{\phi}}_{\boldsymbol{i}}) \prod_{i'\in\mathcal{BG}} p(\boldsymbol{f_{i'}}|\mathcal{B}), \qquad (2)$$

where the foreground $\mathcal{FG}$ is the set of all pixel locations on the feature map $\boldsymbol{F}$ that are covered by the human. The background $\mathcal{BG}$ contains those pixels respectively that are not covered. The foreground likelihood $p(\boldsymbol{f_i}|\hat{\boldsymbol{\phi}}_{\boldsymbol{i}})$ is defined as a Gaussian distribution $\mathcal{N}(\hat{\boldsymbol{\phi}}_{\boldsymbol{i}}, \sigma_i^2\boldsymbol{I})$ with the mean vector $\hat{\boldsymbol{\phi}}_{\boldsymbol{i}}$ at location i, and a standard deviation σ_i. Background features are modeled using a simple background model $p(\boldsymbol{f_{i'}}|\mathcal{B})$ that is defined by a Gaussian distribution $\mathcal{N}(\boldsymbol{b}, \sigma^2\boldsymbol{I})$, where the parameters are $\mathcal{B} = \{\boldsymbol{b}, \sigma\}$ learned from the background features in the training images.

Occlusion robustness. Following related work on occlusion-robust analysis-by-synthesis [9], we define a robust likelihood as:

$$p(\boldsymbol{F}|\mathcal{G}(\boldsymbol{\theta}), \mathcal{B}, \boldsymbol{Z}) = \prod_{i\in\mathcal{FG}} p(\boldsymbol{f_i}|\hat{\boldsymbol{\phi}}_{\boldsymbol{i}}, z_i) \prod_{i'\in\mathcal{BG}} p(\boldsymbol{f_{i'}}|\mathcal{B}) \qquad (3)$$

$$p(\boldsymbol{f_i}|\hat{\boldsymbol{\phi}}_{\boldsymbol{i}}, z_i) = \left[p(\boldsymbol{f_i}|\hat{\boldsymbol{\phi}}_{\boldsymbol{i}})\right]^{z_i} \left[p(\boldsymbol{f_i}|\mathcal{B})\right]^{(1-z_i)},$$

where $z_i \in \{0, 1\}$ is a binary variable and we set its prior probabilities to be $p(z_i{=}1) = p(z_i{=}0) = 0.5$. The variable z_i allows the background model to explain those pixels in $\mathcal{FG}$ that cannot be explained well by the foreground model, presumably due to partial occlusion. To reduce clutter in the remaining paper we will omit the occlusion variable in the coming equations, but note that we are using a robust likelihood during inference.

In the following section, we describe our feature-level generative model for human pose estimation.

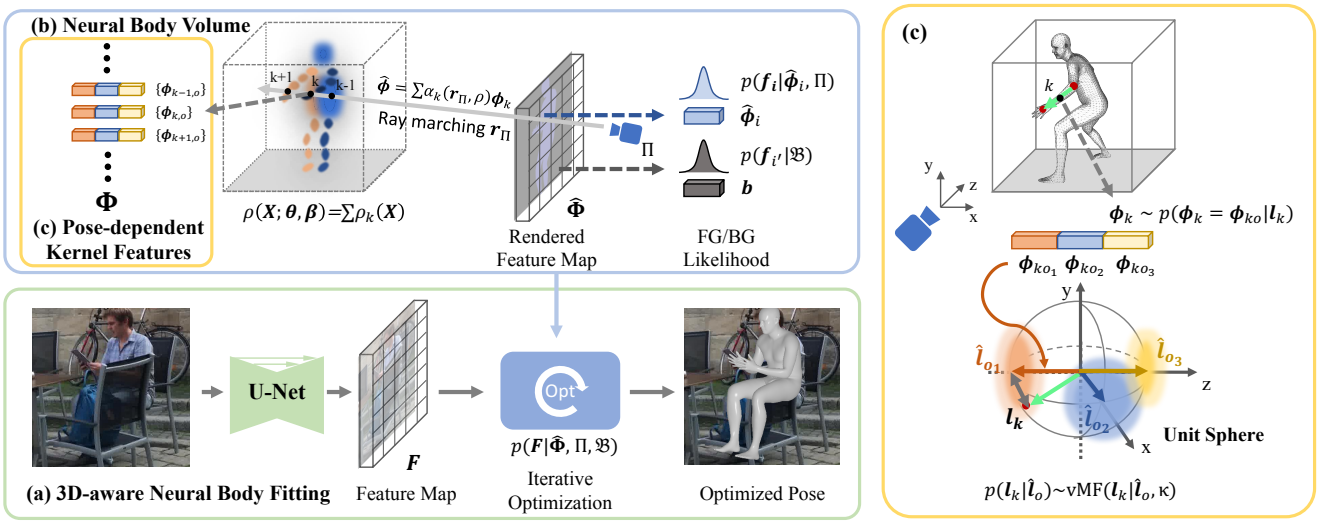

Figure 2: **Overview of our system.** (a) We perform feature-level analysis-by-synthesis for 3D human pose estimation by fitting a 3D-aware generative model of deep feature (NBV) to the feature map $\boldsymbol{F}$ extracted by a U-Net. (b) NBV is defined as a volume representation of human body ρ, driven by pose and shape parameters $\{\boldsymbol{\theta}, \boldsymbol{\beta}\}$, which consists of a set of Gaussian kernels each emitting a pose-dependent feature ϕ. Volume rendering is used to render NBV to a feature map $\hat{\boldsymbol{\Phi}}$. The foreground feature likelihood is defined as a Gaussian distribution centered at the rendered feature vector while the background feature likelihood is modeled by a background model. Pose estimation is done by optimizing the negative log-likelihood (NLL) loss of $\boldsymbol{F}$ w.r.t. $\{\boldsymbol{\theta}, \boldsymbol{\beta}\}$ and camera Π. (c) the distribution of the kernel feature is conditioned on the orientation of the limb that the kernel belongs to.

3.3. Neural Body Volumes

At the core of our framework is the *Neural Body Volumes* (NBV) representation, a model that enables the rendering of human bodies on the feature-level (illustrated in Figure 2(b)). Traditional human body models are mostly mesh-based, e.g. SMPL [39]. However, while meshes are useful representations for forward-rendering applications in computer graphics, they are sub-optimal for differentiable inverse rendering, since the mesh rendering process is inherently difficult to differentiate w.r.t. the model parameters [38]. Prior art [55] showed that volume rendering has a smoother and analytical gradient, and leads to a more efficient optimization, and better handling self-occlusion compared to meshes. Inspired by these results, we propose Neural Body Volumes (NBV), a volume-based representation of human bodies for rendering human bodies on the feature-level. In NBV, a human body is represented by a three-dimensional volume that consists of K Gaussian kernels placed on the body surface. The density at spatial location $\boldsymbol{X} \in \mathbb{R}^3$ is $\rho_k(\boldsymbol{X}) = \mathcal{N}(\boldsymbol{M}_k, \boldsymbol{\Sigma}_k)$. $\boldsymbol{M_k}(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \mathbb{R}^3$ and $\boldsymbol{\Sigma}_k(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \mathbb{R}^{3\times3}$ are the mean vector and covariance matrix conditioned on human pose and shape, parameterized by $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ respectively, controlling the center and shape of the Gaussian kernel which we describe in detail in the following paragraph. The volume density is defined as $\rho(\boldsymbol{X}) = \sum_{k=1}^{K} \boldsymbol{\rho}_k(\boldsymbol{X})$. Each Gaussian kernel is associated with a feature vector $\boldsymbol{\phi_k} \in \mathbb{R}^D$ which can be rendered to image space using volume rendering:

$$\hat{\boldsymbol{\phi}}(\boldsymbol{r}_\Pi) = \int_{t_n}^{t_f} T(t) \sum_{k=1}^{K} \rho_k(\boldsymbol{r}_\Pi(t)) \boldsymbol{\phi_k} \mathrm{d}t, \quad (4)$$
$$\text{where } T(t) = \exp\left(-\int_{t_n}^{t} \rho(\boldsymbol{r}_\Pi(s)) \mathrm{d}s\right),$$

that computes the aggregated feature along the ray $\boldsymbol{r}_\Pi(t)$ $t \in [t_n, t_f]$ from the camera center through a pixel on the image plane where Π denotes the camera parameters. The Gaussian kernel representation enables calculation of the analytic form of the integral $\hat{\boldsymbol{\phi}}(\boldsymbol{r}) = \sum_{k=1}^{K} \alpha_k(\boldsymbol{r}_\Pi, \rho)\boldsymbol{\phi}_k$ which we provide in the supplementary material. Here, the number of Gaussian kernels K and the associated features $\Phi = \{\phi_k\}$ are global parameters shared across all human instances. While for each input image, we optimize the pose $\boldsymbol{\theta}$ and shape $\boldsymbol{\beta}$ to transform the location $\boldsymbol{M_k}(\boldsymbol{\theta}, \boldsymbol{\beta})$ and shape $\boldsymbol{\Sigma}_k(\boldsymbol{\theta}, \boldsymbol{\beta})$ of the Gaussian ellipsoids. Our model can fit arbitrary shapes with a sufficient number of kernels.

Conditioning on pose and shape. Given a set of body joints $\boldsymbol{J} \in \mathbb{R}^{N\times3}$, the pose is defined as their corresponding rotation matrices $\boldsymbol{\Omega} \in \mathbb{R}^{N\times3\times3}$ relative to the template joints $\bar{\boldsymbol{J}}$ in a skeleton tree. We model body articulation using linear blend skinning (LBS) [33] which transforms the center of the Gaussian kernels with transformation linearly blending the accumulated rigid transformations $\boldsymbol{G}(\boldsymbol{J}, \boldsymbol{\Omega}) \in \mathbb{R}^{N\times4\times4}$ of the N body joints (including the

root transformation). And we model body shape variations by displacing the kernels with linear combinations of a set of L basis shape displacements $\boldsymbol{S} \in \mathbb{R}^{L\times K\times 3}$:

$$\boldsymbol{M_k} = \sum_{i=1}^{N} w_{k,i}\boldsymbol{G}_i[\bar{\boldsymbol{M}}_{\boldsymbol{k}} + \sum_{l=1}^{L}\beta_l \boldsymbol{S}_{lk}|\boldsymbol{1}], \tag{5}$$

where $\bar{\boldsymbol{M}}_{\boldsymbol{k}}$ denotes the kernel position in rest pose, and $\sum_{i=1}^{N} w_{k,i} = 1$ and $\sum_{l=1}^{L}\beta_l = 1$ are pose and shape blend weights. $[\cdot|\boldsymbol{1}]$ denotes the homogeneous coordinates. For the spatial covariance, we also perform transformation and blending according to the rotation of the joints:

$$\boldsymbol{\Sigma}_k^{-1} = \sum_{i=1}^{N} w_{k,i}\boldsymbol{R}_i^T\bar{\boldsymbol{\Sigma}}_k^{-1}\boldsymbol{R}_i,\ \boldsymbol{R_i} = \prod_{j\in A(i)}\boldsymbol{\Omega_i}, \tag{6}$$

where $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{k}}$ is the covariance matrix in rest pose, and $A(i)$ is the ordered set ancestors of joint i. This takes into account that the orientation of the Gaussian ellipsoid should rotate with the pose.

The template joints location $\bar{\boldsymbol{J}}$ can also deform according to shape. Specifically, we regress the template joint locations from the locations of the deformed Gaussian kernels $\bar{\boldsymbol{J}} = g(\bar{\boldsymbol{M}} + \sum_{l=1}^{L}\beta_l\boldsymbol{S}_l)$. The common choice for such regressor $g : \mathbb{R}^{K\times 3} \rightarrow \mathbb{R}^{N\times 3}$ is a linear function [39, 50].

In summary, the proposed Neural Body Volume representation enables us to render human bodies on the feature-level using volume rendering process such that for each pixel in the feature map there will be a feature vector $\hat{\boldsymbol{\phi}}$ corresponding to the contribution from all Gaussian kernels.

3.4. A Generative Model of 3D-Aware Features

Related work on feature-level inverse rendering for rigid pose estimation [19, 69] trains the feature extractor ζ such that the features become invariant to changes in the 3D pose. However, for human pose estimation, it is fundamentally important for the feature representation to be 3D-aware, in order to resolve the inherent 2D-3D ambiguity (as shown in Fig. 1). To resolve this problem, we aim to learn *pose-dependent* feature representations that is able to better resolve the 2D-3D ambiguity of human poses.

3D pose-dependent features for NBV. To overcome the 2D-3D ambiguity, we make the generative model 3D-aware. In particular, we impose a distribution on the kernel features $\boldsymbol{\Phi}$ conditioned on the human pose and shape as shown in Fig. 2(c). Therefore, the rendered kernel features explicitly carry 3D pose information. Specifically, we define a set of body limbs $\{(\boldsymbol{J}_i, \boldsymbol{J}_j)|(i,j)\in\mathcal{L}\}$ each defined as an ordered tuple connecting two body joints. The orientation of a limb is defined as $\boldsymbol{l} = (\boldsymbol{J}_j - \boldsymbol{J}_i)/\|\boldsymbol{J}_j - \boldsymbol{J}_i\|$. We first learn to assign each Gaussian kernel in NBV to one limb according to the pose blend weights. Then we associate each kernel with multiple features $\{\boldsymbol{\phi}_o\}$ that correspond to a set of predefined limb orientations $\{\hat{\boldsymbol{l}}_{\boldsymbol{o}} \in \mathbb{R}^3\}_{o=1}^{O}, \|\hat{\boldsymbol{l}}_{\boldsymbol{o}}\| = 1$. The distribution for the feature vector of the Gaussian kernel k is then defined as:

$$p(\boldsymbol{\phi}_k = \boldsymbol{\phi}_{ko}|\boldsymbol{l}_k(\boldsymbol{\theta},\boldsymbol{\beta})) = \frac{p(\boldsymbol{l}_k|\hat{\boldsymbol{l}}_o)}{\sum_{o=1}^{O}p(\boldsymbol{l}_k|\hat{\boldsymbol{l}}_o)}, \tag{7}$$

where $\boldsymbol{l}_k(\boldsymbol{\theta},\boldsymbol{\beta}) \in \mathbb{R}^3$ is the orientation of the limb that Gaussian kernel k belongs to. $p(\boldsymbol{l}_k|\hat{\boldsymbol{l}}_o)$ is the von Mises-Fisher distribution $\mathrm{vMF}(\boldsymbol{l}_k|\hat{\boldsymbol{l}}_o, \kappa_o)$. In the simple case of only one kernel k, the likelihood of feature at foreground pixel i becomes a Gaussian Mixture Model (GMM):

$$p(\boldsymbol{f_i}|\hat{\boldsymbol{\phi}}_{\boldsymbol{i}}) = \sum_{o=1}^{O} p(\boldsymbol{\phi}_k = \boldsymbol{\phi}_{ko}|\boldsymbol{l}_k)\mathcal{N}(\hat{\boldsymbol{\phi}}_{\boldsymbol{ko}}, \sigma_{io}^2\boldsymbol{I}), \tag{8}$$

where $\hat{\boldsymbol{\phi}}_{\boldsymbol{ko}}$ is the feature rendered from $\boldsymbol{\phi}_{ko}$. Intuitively, the rendered feature has different distributions under different 3D limb orientations. Therefore, we can unambiguously infer the 3D pose from the observed features. During inference, we use the expectation of the kernel feature $\mathbb{E}(\boldsymbol{\phi}_k|\boldsymbol{\theta},\boldsymbol{\beta})$ for volume rendering for differentiability.

3.5. Training

Given a set of images $\{\boldsymbol{I}_n\}_{n=1}^{N}$, with ground truth 3D keypoints $\{\hat{\boldsymbol{J}}_n\}_{n=1}^{N}$ and shape $\{\mathcal{V}_n\}_{n=1}^{N}$, we need to learn a set of parameters in NBV: the template Gaussian kernels and the associated features $\{\bar{\boldsymbol{M}}, \bar{\boldsymbol{\Sigma}}, \boldsymbol{\Phi}\}$, the template joints $\bar{\boldsymbol{J}}$, the blend weights $\boldsymbol{W}$, the basis shape displacements $\boldsymbol{S}$ and the joint regressor g. We also need to train the UNet feature extractor ζ. We train our model in separate steps by first learning the pose/shape-related parameters $\{\bar{\boldsymbol{M}}, \bar{\boldsymbol{\Sigma}}, \bar{\boldsymbol{J}}, \boldsymbol{W}, \boldsymbol{S}, g\}$ then the kernel features $\boldsymbol{\Phi}$ and ζ.

Learning pose and shape parameters in NBV. Starting from a downsampled version of a template body mesh model created by artists, we initialize the kernel centers $\bar{\boldsymbol{M}}$ with the locations of the vertices and compute the spatial covariance matrices $\bar{\boldsymbol{\Sigma}}$ based on the distance of the vertices to their neighbors with the desired amount of overlap. Following [39], a manual segmentation of the template mesh is leveraged to obtain the initial template joints $\bar{\boldsymbol{J}}$, the linear joint regressor g, and the blend weights $\boldsymbol{W}$. Then we train all pose-related parameters $\{\bar{\boldsymbol{M}}, \bar{\boldsymbol{J}}, \boldsymbol{W}, g\}$ together with instance-specific pose $\boldsymbol{\theta}$ by minimizing the reconstruction error between the Gaussian kernels and the ground truth shape $\mathcal{V}$. After that, the ground truth shapes are transformed back to the rest pose and the shape basis $\boldsymbol{S}$ is obtained by running PCA on these pose-normalized shapes. We refer the readers to [39] for details as we share a similar training process for this part. Another regressor $\hat{g}$ is trained to regress the ground truth keypoints from kernel centers $\boldsymbol{M}$. In practice, we can directly convert a trained SMPL model [39] to NBV by placing the Gaussian kernels at the vertices on the SMPL mesh.

After training the pose/shape-related parameters, we register our NBV to the train set to obtain the ground truth shape and pose for each training sample $\{\boldsymbol{\theta}_n\}, \{\boldsymbol{\beta}_n\}$. Then, we learn the NBV kernel features and a UNet feature extractor ζ jointly in an iterative manner.

MLE learning of NBV kernel features. If ζ is trained, we can learn the kernel features $\boldsymbol{\Phi}$ through maximum likelihood estimation (MLE) by minimizing the following negative log-likelihood of the feature representations over the whole training set,

$$\mathcal{L}_{\text{NLL}}(\hat{\boldsymbol{F}}, \boldsymbol{\Phi}) = -\sum_{i \in \mathcal{FG}} \log p(\hat{\boldsymbol{f}}_i | \hat{\boldsymbol{\phi}}_i), \tag{9}$$

where for training efficiency, we use an approximate solution to avoid matrix inversion $\boldsymbol{\phi}_{ko} = \frac{\sum_{i \in \mathcal{K}} \gamma_{iko} \hat{\boldsymbol{f}}_i}{\sum_{i \in \mathcal{K}} \gamma_{iko}}$, where $\mathcal{K}$ is the set of pixels in the training data that the kernel feature $\boldsymbol{\phi}_{ko}$ contributes to, and γ_{iko} is the contribution of $\boldsymbol{\phi}_{ko}$ to pixel i which is obtained from the volume rendering process. Similarly, the parameters of the background distribution are learned using MLE on the features that are not covered by the projected NBV model in the training data. To reduce the computational cost, we follow [69] and employ a momentum strategy [13] to update $\boldsymbol{\Phi}$ in a moving average manner.

3D-aware contrastive learning of the UNet feature extractor. Given the generative model, we can train the UNet feature extractor with the NLL loss as defined in Equation 9 w.r.t. the network parameters. In addition, we want the extracted feature map to have the property that the rendered feature from NBV in the ground truth pose has the largest probability. To this end, we incorporate a set of contrastive losses:

$$\mathcal{L}_{\text{FG}}(\boldsymbol{F}, \mathcal{FG}) = -\sum_{i \in \mathcal{FG}} \sum_{i' \in \mathcal{FG} \setminus \{i\}} \|\boldsymbol{f}_i - \boldsymbol{f}_{i'}\|^2 \tag{10}$$

$$\mathcal{L}_{\text{3D}}(\boldsymbol{F}) = -\sum_{k} \sum_{o} \sum_{o' \in \mathcal{O} \setminus \{o\}} \sum_{i \in \mathcal{K}_{ko}} \sum_{j \in \mathcal{K}_{ko'}} \|\boldsymbol{f}_{iko} - \boldsymbol{f}_{jko'}\|^2 \tag{11}$$

$$\mathcal{L}_{\text{BG}}(\boldsymbol{F}, \mathcal{FG}, \mathcal{BG}) = -\sum_{i \in \mathcal{FG}} \sum_{j \in \mathcal{BG}} \|\boldsymbol{f}_i - \boldsymbol{f}_j\|^2 \tag{12}$$

where $\mathcal{L}_{\text{FG}}$ encourages features of different pixels to be distinct from each other. $\mathcal{L}_{\text{3D}}$ encourages features of the same kernel in different 3D poses to be distinct from each other, i.e. to become 3D-aware. $\mathcal{L}_{\text{BG}}$ encourages features on the human to be distinct from those in the background. We optimize those losses jointly $\mathcal{L}_{\text{contrast}} = \mathcal{L}_{\text{FG}} + \mathcal{L}_{\text{3D}} + \mathcal{L}_{\text{BG}}$ in a contrastive learning framework. Therefore, the total loss for training ζ is $\mathcal{L}_{train} = \mathcal{L}_{\text{NLL}} + \mathcal{L}_{\text{contrast}}$.

Bottom-up initialization with regression heads. For efficient inference, it is a common practice in generative modeling to initialize with regression-based methods. In our model, we add a regression head to the UNet feature extractor to predict the pose and shape parameters $\{\boldsymbol{\theta}, \boldsymbol{\beta}\}$ from the observed feature map. The regression head and the UNet are learned jointly.

3.6. Inference

We estimate the 3D human pose $\boldsymbol{\theta}$, shape parameters $\boldsymbol{\beta}$, and the camera parameters Π using the analysis-by-synthesis formulation in Equation 1. This boils down to minimizing an NLL loss plus a regularization term from the pose prior $p(\boldsymbol{\theta})$ w.r.t $\{\boldsymbol{\theta}, \boldsymbol{\beta}, \Pi\}$. The initialization comes from the regression head. Our proposed generative model is fully differentiable and therefore can be optimized using gradient-based methods.

3.7. Implementation Details

We convert the neutral SMPL model to NBV using the method described in Sec. 3.5, keeping 858 kernels. We use a U-Net [58] style network as the feature extractor which consists of a ResNet-50 [14] backbone and 3 upsampling blocks. The regression head follows the design of [25]. The input image is a 320×320 crop centered around the human. The feature map has a $4\times$ downsampled resolution and the feature dimension is 64 which balances performance and computation cost as shown in ablation in Sec. 4.3. The Adam optimizer with a learning rate of 5×10^{-5} and batch size of 64 is used for training the feature extractor and the regression head. Standard data augmentation techniques are used including random flipping, scaling, and rotation. For the 3D pose-dependent features, we consider the limb orientation projected to the yz-plane and split the unit circle evenly. We set O=4 for all kernels which already gives good enough results as shown in the ablation study in Sec. 4.3. We consider 9 limbs including the left/right upper/lower arm/leg and the torso. The torso includes the head and its orientation is defined as the direction from the mid-hip joint to the neck joint. For inference, we also use Adam as the optimizer with a learning rate of 0.02 and run a maximum of 80 steps. We use VPoser [50] as our 3D pose prior. We check the negative log-likelihood $\mathcal{L}_{NLL}$ of the initial pose and its 180°-rotated version around y-axis and use the better one to initialize our model. Inference speed is $\sim$ 1.7fps with a batch size of 32 on 4 NVIDIA Titan Xp GPUs.

4. Experiments

In this section, we demonstrate the effectiveness and robustness of 3DNBF by comparing it with SOTA HPE. In addition to existing benchmarks, we propose a more challenging adversarial evaluation for occlusion robustness. Finally, we conduct ablation studies to verify the design choices and effectiveness of different components.

4.1. Training Setup and Datasets

Training. Follow the common setting, we train NBV on `Human3.6M` [18], `MPI-INF-3DHP` [40], and

Method	3DPW [68]			3DPW-Occ [78]			3DPW-AdvOcc@40			3DPW-AdvOcc@80			3DOH50K [78]		
	MPJPE↓	P-MPJPE↓	PCKh↑	MPJPE↓	P-MPJPE↓	PCKh↑	MPJPE↓	P-MPJPE↓	PCKh↑	MPJPE↓	P-MPJPE↓	PCKh↑	MPJPE↓	P-MPJPE↓	PCKh↑
SPIN [27]	96.6	58.3	91.7	97.5	60.8	85.9	203.5	97.0	63.6	338.8	111.7	34.1	101.3	67.9	83.3
HMR-EFT [21]	89.5	53.4	93.1	95.8	57.1	87.2	146.7	73.2	77.8	202.8	83.7	63.7	97.4	65.8	84.4
MGraphr [34]	80.4	53.4	88.7	116.8	75.7	66.6	158.8	93.2	70.8	261.5	121.0	48.8	127.4	76.0	79.8
PARE [25]	81.4	50.9	92.5	86.8	58.8	86.2	126.5	72.5	82.3	210.9	97.4	61.9	100.7	65.1	84.2
3DNBF	**79.8**	**49.3**	**95.7**	**77.2**	**51.2**	**93.1**	**105.1**	**60.5**	**92.0**	**140.7**	**71.8**	**85.0**	**86.7**	**57.5**	**88.6**

Table 1: **Evaluation on 3DPW, 3DPW-Occ, 3DPW-AdvOcc, and 3DOH50K.** The number 40 and 80 after 3DPW-AdvOcc denote the occluder size. Note the performance improvement of 3DNBF increases as occlusion becomes more severe. (P-MPJPE: PA-MPJPE; MGraphr: Mesh Graphormer.)

Method	3DPW [68]			3DPW-AdvOcc@40			3DPW-AdvOcc@80		
	MPJPE↓	PA-MPJPE↓	PCKh↑	MPJPE↓	PA-MPJPE↓	PCKh↑	MPJPE↓	PA-MPJPE↓	PCKh↑
HMR-EFT [21]	89.5	53.4	93.1	146.7	73.2	77.8	202.8	83.7	63.7
+ SMPLify [4]	106.2	64.8	91.2	133.4	75.8	85.6	192.2	89.3	73.4
+ 3DPOF [73]	97.6	60.8	90.6	125.1	69.6	84.0	175.0	78.8	73.0
+ EFT [21]	92.8	55.9	93.0	114.1	64.6	88.7	158.2	75.2	78.5
+ 3DNBF	**88.8**	**53.3**	**93.6**	**109.4**	**62.2**	**90.9**	**150.2**	**72.0**	**85.3**

Table 2: **Comparison to optimization-base methods.** HMR-EFT is used for initialization.

COCO [35] datasets. We use ground truth SMPL fittings for Human3.6M and MPI-INF-3DHP [23, 27] and the pseudo-ground truth fittings from EFT [21] for COCO following [25]. The selection of subjects for training strictly follows previous work [21,25,27]. We first train the feature extractor on COCO for 175K iterations, then fine-tune on all data for another 175K iterations. During fine-tuning, the sampling ratio in each batch is 50% Human3.6M, 20% MPI-INF-3DHP and 30% COCO. Note that for all baseline methods, we use the official model trained with the *same* data as ours for fairness.

Occlusion Robustness Evaluation. We conduct evaluations on two datasets to measure the robustness and generalization of our method: an in-the-wild dataset 3DPW-Occ [78] which is a subset of the original 3DPW [68] dataset and an artificial indoor occlusion dataset 3DOH50K [78]. In particular, we directly test all models on these datasets *without any training on them*. For 3DPW and 3DPW-Occ, we sample the videos every 30 frames. We report mean per joint position error (MPJPE) and Procrustes-aligned mean per joint position error (PA-MPJPE) in mm as the main evaluation metrics. We also report the 2D Percentage of Correct Keypoints with head length threshold (PCKh) to measure how well the prediction aligns with the 2D image.

Adversarial occlusion robustness evaluation. Inspired by the occlusion analysis in [25], we design an adversarial protocol 3DPW-AdvOcc to further evaluate the occlusion robustness of SOTA methods. Specifically, we slide an occlusion patch over the input image to find the worst prediction. This is done by comparing the relative performance degradation on the visible joints. We argue that evaluating the performance on occluded joints is sometimes ambiguous since the location of occluded joints is not always predictable even for human. Therefore, for a more stable and meaningful evaluation, the joints outside the bounding box or occluded by the patch are *excluded from evaluation*. Instead of using a gray occlusion patch, we use textured patches generated by randomly cropping texture maps from the Describable Textures Dataset (DTD) [8], which is more challenging. Two different square patch sizes are used: 40 and 80 relative to a 224×224 image, denoted as Occ@40 and Occ@80 respectively, and the stride is set to 10.

4.2. Performance Evaluation

Baselines. To demonstrate the superior performance and occlusion robustness of 3DNBF, we compare our model with four SOTA regression-based methods: SPIN [27], HMR-EFT [21], Mesh Graphormer [34] and PARE [25] where PARE is designed to be robust to occlusions with part attention and trained with synthetic occlusion augmentation. For fair comparisons, we adopt the models with the *same* ResNet-50 backbone for all methods. We also compare 3DNBF with SOTA optimization-base methods that also improve occlusion robustness: SMPLify [4], 3D POF [73] and EFT [21].

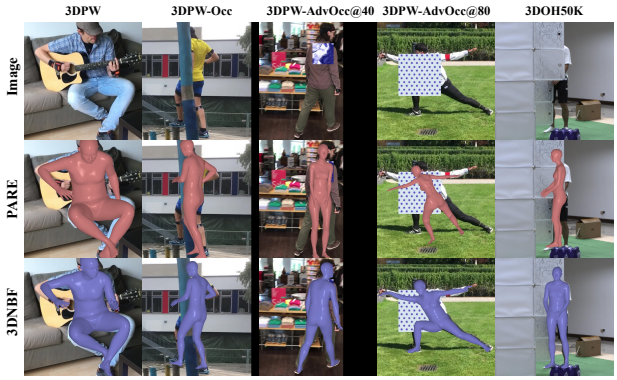

Figure 3: **Qualitative results on evaluated datasets.**

Comparison to SOTA. As shown in Table 1, we first evaluate on the standard `3DPW` test set and 3DNBF achieves SOTA performance. On occlusion datasets `3DPW-Occ` and `3DOH50K`, our improvement becomes more significant. We then evaluate the occlusion robustness on `3DPW-AdvOcc` where we find all regression-based methods suffer from occlusion with the MPJPE increasing up to 225% and the PA-MPJPE increasing up to 127% even for the best-performing method. The transformer-based model [34] suffers the most from occlusion which we speculate to be due to overfitting. In contrast, 3DNBF is much more robust to occlusion improving over the SOTAs by a wide margin. Note that our predictions align better with the image as shown in PCKh.

Comparison to other optimization-based methods. We compare 3DNBF with three optimization-based methods on `3DPW` and `3DPW-AdvOcc`. We choose HMR-EFT as initial regressor to use the official EFT implementation. All methods use the same 2D keypoints detected by OpenPose [5]. As shown in Table 2, we achieve the best performance on both non-occluded and occluded settings. Although SMPLify improves 2D PCKh, it does not quite improve on the 3D metrics. This is due to SMPLify only fitting SMPL parameters to 2D keypoints without capturing any 3D information from the image, thus suffering from the 2D-3D ambiguity. EFT fine-tunes the regression network using a 2D keypoint reprojection loss. It achieves better performance than SMPLify because the regression network itself can implicitly encode 3D information in the input image and serve as a conditional 3D pose prior. However, EFT does not improve on non-occluded cases.

Qualitative results. We qualitatively demonstrate the improved occlusion robustness of 3DNBF compared to the SOTA regression-based method PARE in Fig. 3 and show more comparisons in the supplementary material. Notice that PARE makes unaligned predictions for visible joints.

4.3. Ablation Studies

In this section, we provide ablation of different components in 3DNBF including the NBV, the design of pose-dependent features, and the 3D-aware contrastive loss. All experiments are on `3DPW-AdvOcc@80`.

	MPJPE↓	P-MPJPE↓	PCKh↑
3DNBF	**140.7**	**71.8**	**85.0**
Init. only	171.4	80.5	75.8
w/o NBV	146.6	72.6	83.1
w/o contrast	167.4	80.4	79.4

(a) Ablations for 3DNBF.

O	MPJPE↓	P-MPJPE↓	PCKh↑
1	189.6	84.8	69.2
4	**140.7**	**71.8**	**85.0**
8	161.2	77.8	81.0

(b) # of pose-dependent features. O=1 means using pose-independent feature.

Table 3: **Ablation studies.** All experiments are performed on `3DPW-AdvOcc@80`. (P-MPJPE: PA-MPJPE.)

NBV vs. Mesh. To demonstrate the advantage of NBV over mesh representation, we replace it with a mesh-based neural representation using SMPL while keeping everything else the same. We use the differentiable rendering implementation from PyTorch3D [54]. As shown in Table 3a, this model achieves worse results than using NBV.

The pose-dependent kernel features. The 3D-aware pose-dependent kernel feature is key to the success of 3DNBF. Here we validate its effectiveness by comparing it with pose-independent feature (O=1). As shown in Table 3b, much better performance is achieved with O>1. Using 4 features for each kernel achieves the best performance while further increasing the number of features may make the learning harder as it introduced more parameters.

Importance of contrastive training. We ablate this by training our model without contrastive learning, i.e. training the feature extractor with regression loss only. The performance degrades a lot as shown in Table 3a. The intuition is that our model requires the features to be Gaussian distributed and contrastive learning encourages this.

Regression head performance. Although we do not expect the regression head to be robust to occlusion, it achieves higher occlusion robustness compared to other regression-based methods as shown in Table 3a and Table 1.

5. Conclusion

In this work, we introduce 3D Neural Body Fitting (3DNBF) - an approximate analysis-by-synthesis approach to 3D HPE that is accurate and highly robust to occlusion. To this end, NBV is proposed which is an explicit volume-based generative model of pose-dependent features for human body. We propose a contrastive learning framework for training a feature extractor that captures the 3D pose information of the body parts thus overcoming the 2D-3D ambiguity in monocular 3D HPE. Experiments on challenging benchmark datasets demonstrate that 3DNBF outperforms SOTA regression-based methods as well as optimization-based methods. While focusing on occlusion robustness in this paper, we expect our model to be robust to other challenging adversarial examinations [59, 62].

6. Acknowledgements

This work was supported by NIH R01 EY029700. This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via [2022-21102100005]. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The US. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

[1] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: Shape completion and animation of people. *SIGGRAPH*, 2005. 2

[2] Yutong Bai, Angtian Wang, Adam Kortylewski, and Alan Yuille. Coke: Contrastive learning for robust keypoint detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 65–74, 2023. 2

[3] Benjamin Biggs, David Novotny, Sebastien Ehrhardt, Hanbyul Joo, Ben Graham, and Andrea Vedaldi. 3d multi-bodies: Fitting sets of plausible 3d human models to ambiguous image data. In *Advances in Neural Information Processing*, 2020. 2

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it SMPL: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision*, pages 561–578. Springer, 2016. 1, 2, 3, 7

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017. 8

[6] Yu Cheng, Bo Yang, Bo Wang, and Robby T Tan. 3d human pose estimation using spatio-temporal networks with explicit occlusion training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10631–10638, 2020. 2

[7] Yu Cheng, Bo Yang, Bo Wang, Wending Yan, and Robby T Tan. Occlusion-aware networks for 3d human pose estimation in video. In *International Conference on Computer Vision*, pages 723–732, 2019. 2

[8] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 7

[9] Bernhard Egger, Sandro Schönborn, Andreas Schneider, Adam Kortylewski, Andreas Morel-Forster, Clemens Blumer, and Thomas Vetter. Occlusion-aware 3d morphable models and an illumination prior for face image analysis. *International Journal of Computer Vision*, 126(12):1269–1287, 2018. 3

[10] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Košecká, and Ziyan Wu. Hierarchical kinematic human mesh recovery. In *European Conference on Computer Vision*, pages 768–784. Springer, 2020. 2

[11] Riza Alp Guler and Iasonas Kokkinos. HoloPose: Holistic 3D human reconstruction in-the-wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10884–10894, 2019. 1, 2

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 2

[13] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 6

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[15] Buzhen Huang, Yuan Shu, Jingyi Ju, and Yangang Wang. Occluded human body capture with self-supervised spatial-temporal motion prior. *arXiv preprint arXiv:2207.05375*, 2022. 2

[16] Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V. Gehler, Javier Romero, Ijaz Akhter, and Michael J. Black. Towards accurate marker-less human shape and pose estimation over time. In *International Conference on 3DVision*, 2017. 2

[17] Peter J Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004. 2

[18] 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, 2013. 6

[19] Shun Iwase, Xingyu Liu, Rawal Khirodkar, Rio Yokota, and Kris M Kitani. Repose: Fast 6d object pose refinement via deep texture rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3303–3312, 2021. 5

[20] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *European Conference on Computer Vision*, 2022. 2, 3

[21] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *2021 International Conference on 3D Vision (3DV)*, pages 42–52. IEEE, 2021. 1, 2, 7

[22] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8320–8329, 2018. 2

[23] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and

pose. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 1, 2, 7

[24] Daniel Kersten, Pascal Mamassian, and Alan Yuille. Object perception as bayesian inference. *Annu. Rev. Psychol.*, 55:271–304, 2004. 1

[25] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11127–11137, 2021. 1, 2, 6, 7

[26] 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 *International Conference on Computer Vision*, pages 2252–2261, 2019. 1, 2

[27] 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 *International Conference on Computer Vision*, pages 2252–2261, 2019. 1, 2, 7

[28] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11605–11614, 2021. 2

[29] Adam Kortylewski, Ju He, Qing Liu, and Alan L Yuille. Compositional convolutional neural networks: A deep architecture with innate robustness to partial occlusion. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8940–8949, 2020. 3

[30] Adam Kortylewski, Qing Liu, Angtian Wang, Yihong Sun, and Alan Yuille. Compositional convolutional neural networks: A robust and interpretable model for object recognition under occlusion. *International Journal of Computer Vision*, pages 1–25, 2020. 1, 3

[31] Adam Kortylewski, Qing Liu, Huiyu Wang, Zhishuai Zhang, and Alan Yuille. Combining compositional models and deep networks for robust object classification under occlusion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1333–1341, 2020. 3

[32] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4704–4713, 2017. 2

[33] John P Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 165–172, 2000. 4

[34] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12939–12948, 2021. 7, 8

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 7

[36] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Transactions on Graphics (TOG)*, 40(6):1–16, 2021. 2, 3

[37] Qihao Liu, Yi Zhang, Song Bai, and Alan Yuille. Explicit occlusion reasoning for multi-person 3d human pose estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 497–517. Springer, 2022. 2

[38] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7708–7717, 2019. 4

[39] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. In *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 2015. 1, 2, 4, 5

[40] 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 *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017. 2, 6

[41] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. Xnect: Real-time multi-person 3d motion capture with a single rgb camera. *Acm Transactions On Graphics (TOG)*, 39(4):82–1, 2020. 2

[42] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*, pages 120–130. IEEE, 2018. 2

[43] Pol Moreno, Christopher KI Williams, Charlie Nash, and Pushmeet Kohli. Overcoming occlusion with inverse graphics. In *European Conference on Computer Vision*, pages 170–185. Springer, 2016. 2

[44] Natalia Neverova, David Novotny, Marc Szafraniec, Vasil Khalidov, Patrick Labatut, and Andrea Vedaldi. Continuous surface embeddings. *Advances in Neural Information Processing Systems*, 33:17258–17270, 2020. 3

[45] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7588–7597, 2019. 3

[46] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *CVPR*, 2021. 3

[47] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *International Conference on Computer Vision*, pages 5762–5772, 2021. 2

[48] 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. In *International Conference on 3DVision*, 2018. 1, 2

[49] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5865–5874, 2021. 2

[50] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019. 1, 2, 3, 5, 6

[51] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[52] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 459–468, 2018. 1, 2

[53] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 2, 3

[54] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv preprint arXiv:2007.08501*, 2020. 8

[55] Helge Rhodin, Nadia Robertini, Christian Richardt, Hans-Peter Seidel, and Christian Theobalt. A versatile scene model with differentiable visibility applied to generative pose estimation. In *International Conference on Computer Vision*, pages 765–773, 2015. 2, 3, 4

[56] Chris Rockwell and David F. Fouhey. Full-body awareness from partial observations. In *European Conference on Computer Vision*, pages 522–539, 2020. 2

[57] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3433–3441, 2017. 2

[58] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 6

[59] Nataniel Ruiz, Adam Kortylewski, Weichao Qiu, Cihang Xie, Sarah Adel Bargal, Alan Yuille, and Stan Sclaroff. Simulated adversarial testing of face recognition models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4145–4155, 2022. 8

[60] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *International Conference on Computer Vision*, pages 2304–2314, 2019. 2

[61] István Sárándi, Timm Linder, Kai O Arras, and Bastian Leibe. How robust is 3d human pose estimation to occlusion? *arXiv preprint arXiv:1808.09316*, 2018. 2

[62] Michelle Shu, Chenxi Liu, Weichao Qiu, and Alan Yuille. Identifying model weakness with adversarial examiner. In *AAAI*, volume 34, pages 11998–12006, 2020. 8

[63] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. *Advances in Neural Information Processing Systems*, 34:12278–12291, 2021. 2, 3

[64] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *European Conference on Computer Vision*, pages 529–545, 2018. 2

[65] Jun Kai Vince Tan, Ignas Budvytis, and Roberto Cipolla. Indirect deep structured learning for 3D human shape and pose prediction. In *British Machine Vision Conference*, 2017. 1, 2

[66] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 3

[67] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *Advances in Neural Information Processing*, pages 5236–5246, 2017. 1, 2

[68] 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 *European Conference on Computer Vision*, pages 601–617, 2018. 2, 7

[69] Angtian Wang, Adam Kortylewski, and Alan Yuille. Nemo: Neural mesh models of contrastive features for robust 3d pose estimation. *arXiv preprint arXiv:2101.12378*, 2021. 1, 2, 3, 5, 6

[70] Justin Wang, Edward Xu, Kangrui Xue, and Lukasz Kidzinski. 3D pose detection in videos: Focusing on occlusion. *arXiv preprint arXiv:2006.13517*, 2020. 2

[71] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 16210–16220, 2022. 2, 3

[72] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 2

[73] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10965–10974, 2019. 1, 2, 7

[74] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 13286–13296. IEEE, 2022. 2

[75] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. Denserac: Joint 3d pose and shape estimation by dense render-and-compare.

In *International Conference on Computer Vision*, October 2019. 1, 2

[76] Alan Yuille and Daniel Kersten. Vision as bayesian inference: analysis by synthesis? *Trends in cognitive sciences*, 10(7):301–308, 2006. 1

[77] 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*, pages 465–481, 2020. 1

[78] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7374–7383, 2020. 2, 7

[79] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *International Conference on Computer Vision*, pages 398–407, 2017. 2