D-IF: Uncertainty-aware Human Digitization via Implicit Distribution Field

Xueting Yang[1†] Yihao Luo[1,2†] Yuliang Xiu[3] Wei Wang[1] Hao Xu[1,4] Zhaoxin Fan[1,4*]
[1]Psyche AI Inc. [2]Imperial College London [3]Max Planck Institute for Intelligent Systems
[4]Hong Kong University of Science and Technology
yangxueting@psyai.net y.luo23@imperial.ac.uk yuliang.xiu@tue.mpg.de
faithwwei@bupt.edu.cn {hxubl,zfanaq}@connect.ust.hk

Abstract

Realistic virtual humans play a crucial role in numerous industries, such as metaverse, intelligent healthcare, and self-driving simulation. But creating them on a large scale with high levels of realism remains a challenge. The utilization of deep implicit function sparks a new era of image-based 3D clothed human reconstruction, enabling pixel-aligned shape recovery with fine details. Subsequently, the vast majority of works locate the surface by regressing the deterministic implicit value for each point. However, should all points be treated equally regardless of their proximity to the surface? In this paper, we propose replacing the ***implicit value*** *with an* ***adaptive uncertainty distribution****, to differentiate between points based on their distance to the surface. This simple "value ⇒ distribution" transition yields significant improvements on nearly all the baselines. Furthermore, qualitative results demonstrate that the models trained using our uncertainty distribution loss, can capture more intricate wrinkles, and realistic limbs. Code and models are available for research purposes at* github.com/psyai-net/D-IF_release.

1. Introduction

The creation of realistic digital avatars with intricate clothing details holds significant importance in various applications, such as metaverse [40], intelligent healthcare [45], teleportation [19], and self-driving simulation [44]. However, conventional methods require substantial human resources and considerable costs for designing or capturing high-fidelity 3D digital avatars. To simplify this process while maintaining quality, both the academic community and industry have shifted their focus and efforts toward data-driven approaches [20, 37, 43], to accurately reconstruct 3D humans from images or monocular videos.

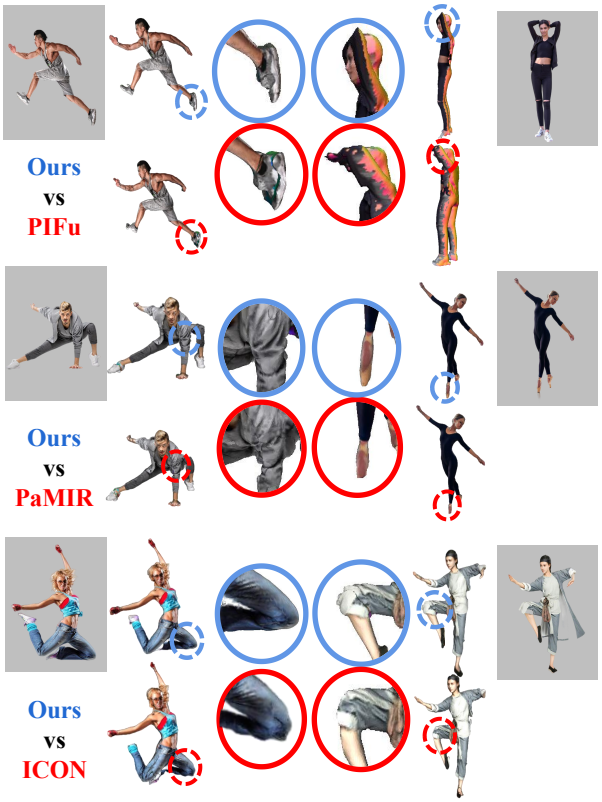

Figure 1. **Comparison with mainstream SOTAs**. Unlike PaMIR [47], PIFu [37], and ICON [43], which frequently produce 3D humans with distorted or non-human shape limbs, missing details, and high-frequency noise, our method overcomes these issues and achieves superior geometric details in reconstruction.

Explicit-shape-based approaches [1–3, 18, 41, 48] typically fit and deform the parametric body model, like SMPL-(X) [23, 33], to align with the visual observations. While these approaches could recover 3D humans wearing tight clothing, they face limitations when it comes to reconstructing loose-fitting garments that largely deviate from the body. Alternatively, implicit-function-based approaches [4, 13–15, 21, 38, 42, 43, 47] utilize implicit function parame-

[†]These authors contributed equally to this work.
[*]represents the corresponding author.

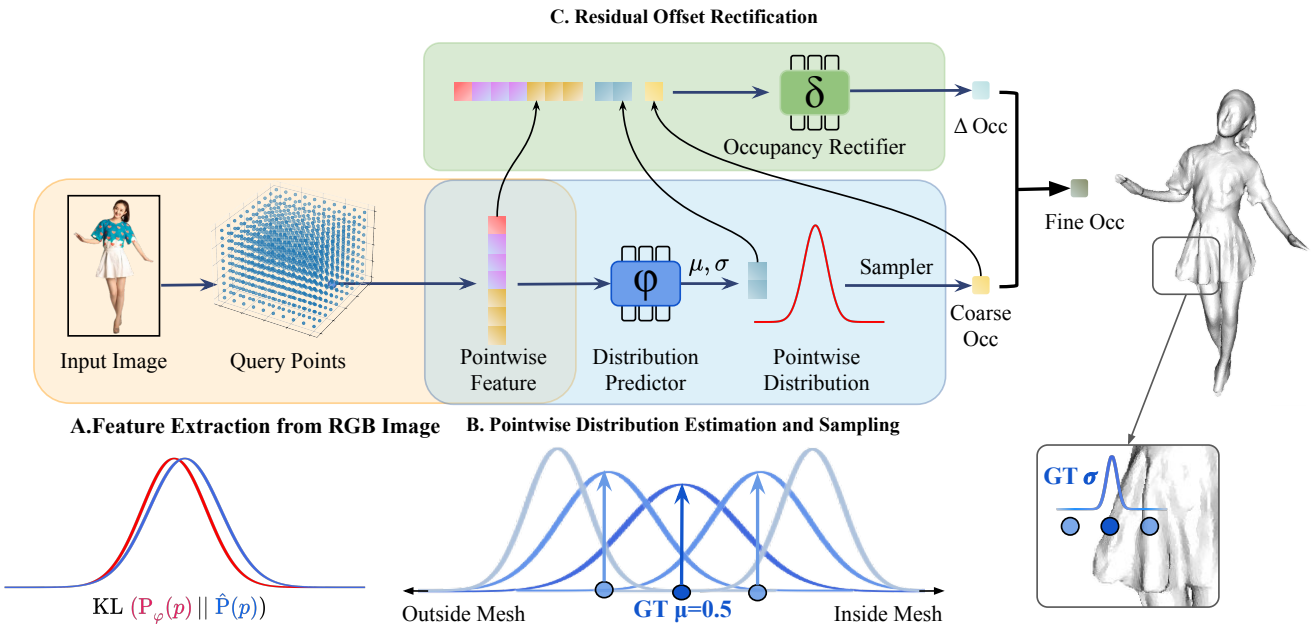

Figure 2. **The framework of D-IF (Sec. 3.1). A.** For queried points $p \in \mathbb{R}^3$, which are uniformly sampled from the entire 3D space, 7D pointwise local feature $F_{7\mathrm{D}}$ is extracted. **B.** $F_{7\mathrm{D}}$ is then fed into distribution predictor $P_\varphi\,(p)$, to estimate the per-point distribution (μ, σ), and the coarse occupancy $O_s(p)$ is sampled from it. **C.** Given all above features and outputs (9D), Occupancy Rectifier, an additional MLP, is to predict the residual offset $\Delta O_s(p)$. Finally, the fine occupancy $\widetilde{O_s}(p)$ is obtained via $\widetilde{O_s}(p) = O_s(p) + \Delta O_s(p)$. The bottom half of this illustration demonstrates the design of uncertainty-aware supervised learning (Sec. 3.4). During training, we formulate the "pseudo" ground-truth distribution $\hat{P}\,(p)$ as follows: ground-truth smooth occupancy value $O_{gt}(p)$ as the expectation μ (Eq. (5)), and σ is gradually reducing as the point-to-mesh distance increases (Eq. (6)). Finally, KL-divergence loss is introduced to minimize the difference between predicted and pseudo distributions, see Eq. (7).

terized by MLPs to regress occupancy fields [28] or signed distance fields (SDF) [32]. Detailed meshes can then be extracted using Marching Cubes [24] from the iso-surface of a certain implicit value. Implicit methods have demonstrated their superiority in capturing geometric details and accommodating various topological structures. However, they may generate non-human shapes for unseen poses or garments, due to the absence of shape regularization.

Despite the impressive results of prior implicit-based methods, they have not fully taken into account the presence of *"uncertainty"* in the geometric deformation that arises during the reconstruction procedure. For example, in the case of points that significantly deviate from the body, a practical shortcut would be to categorize them as "outside points", while even a minor estimation disturbance near the surface can lead to a completely wrong occupancy result.

To account for such uncertainty, this paper introduces implicit distribution fields, called D-IF. Instead of directly estimating the implicit value at each point, we opt to sample the implicit value from an estimated distribution. This enables us to effectively capture the uncertainty associated with the distance from the surface. The overview of our method is shown in Fig. 2. Inspired by ICON [43], we firstly extract 7D local features from the input image and estimated SMPL body. These features are then utilized to estimate the point-wise occupancy distribution. Sampling from the projected distribution of grid points across the entire 3D space produces a coarse occupancy field. To enhance the level of detail, we introduce an additional MLP called "Occupancy Rectifier", which refines the coarse occupancy field further, to obtain the fine occupancy field, the final clothed mesh is extracted using Marching Cubes [24] at 0.5 level-set.

Upon delving deeper into D-IF, it becomes apparent that there exists a dilemma for the learned distribution to be simultaneously accurate and uncertain. This necessitates finding a balance in terms of distribution sharpness. To address this dilemma, we introduce an explicit supervision mechanism known as the uncertainty distribution loss to learn the distribution, which is illustrated in Fig. 2. The insight behind the loss is based on the assumption that pointwise distribution is highly relevant to point-to-mesh distance. We elaborate this in Sec. 3.4, where we introduce the KL-divergence [22] between the predicted distribution and a designed distribution (pseudo GT). Moreover, the Occupancy Rectifier module aims to correct any erroneous occupancy while simultaneously refining intricate shape details.

Quantitative experiments on CAPE [26] confirm that our method achieves SOTA performance, see Tab. 1. And as a "plug-and-play" module, it significantly improves the reconstruction accuracy on nearly all the mainstream baselines. As depicted in Fig. 1, D-IF excels at recovering intricate geometric features, while mitigating common artifacts found in other works [37, 43, 47], including non-human parts, missing details, and high-frequency noise.

2. Related Work

2.1. Explicit-shape based human reconstruction

Parametric models have been widely used in 3D human reconstruction. Previous works [1, 3, 18] have introduced the concept of SMPL+D, where displacement is added to the vertices of the SMPL [23] model to represent clothed models. For example, Tex2Shape [3] defines the vertex offset in the UV space of SMPL to achieve higher-resolution representations with detailed clothing wrinkles. While MGN [6] performs vertex segmentation on SMPL for different clothing types, enabling better expression of clothing boundaries in reconstructed SMPL+D representations. Alternative parametric methods, inspired by the representation of SMPL+D, propose vertex deformations on parametric SMPL models to capture more geometric details. For instance, Zhu et al. [48] employ hierarchical free-form 3D deformation to improve the geometry of the predicted human body. While Weng et al. [39] propose using normal directions to improve deformations and obtain better clothed human body meshes from the SMPL model. Though these methods achieve acceptable performance, there are limitations regarding their ability to express various clothing types due to inherent topology constraints imposed by parametric models. Additionally, learning geometric details from explicit parametric models can be challenging.

2.2. Implicit-based human reconstruction

Implicit representations are employed in clothed human reconstruction [4, 9, 13, 14, 16, 25, 42, 43] to overcome the constraints associated with parametric representations, which have emerged as the prevailing methods. Among these methods, PIFu [37] is the first to employ pixel-aligned features to regress the occupancy field. PIFuHD [38] utilizes a coarse and fine network structure and additional normals as additional geometric information to improve PIFu. PaMIR [47] employs SMPL shape prior as shape regularization. ICON [43] and ECON [42] still utilize the SMPL-(X) body prior but focus on improving the pose robustness and garment topological flexibility respectively. In contrast to the above-mentioned methods, which are limited to static mesh outputs which are not ready for animation, other methods such as ARCH [16], ARCH++ [14], aim to reconstruct 3D humans in canonical space.

While the methods mentioned above have demonstrated satisfactory outcomes in cloth reconstruction, they depend on deterministic predictions of implicit values for each point and disregard the significance of incorporating uncertainties inherent in the cloth reconstruction process.

2.3. Distribution-based 3D reconstruction

In recent years, research on distribution-based implicit functions has drawn attention. For instance, MaGNet [5] estimates a probability distribution of single-view depth, achieving higher accuracy yet evaluating fewer depth candidates. This approach is similar to the one proposed in [36]. SubFocal [8] estimates the Dirac delta distribution within the range of pixel depth for the input image and uses it as a backpropagation supervisor to reduce bad-pixels. Additionally, CaDDN [34] projects rich contextual features into the appropriate depth interval in 3D space using a projected categorical depth distribution for each pixel. Distribution-based methods have also been applied to point cloud completion and pose estimation tasks, where they learn distributions for shape completion or motion changes [31, 35].

Although the above methods have achieved significant improvements, most existing approaches tend to overlook the variations in distribution among different spatial points. In this study, we propose not only a distribution to express the uncertainty of clothing but also a method to differentiate between near-surface and floating points through the utilization of the proposed uncertainty distribution loss. To our knowledge, this study represents the first attempt to address the uncertainties related to clothed human reconstruction.

3. Method

Our study aims to generate detailed human meshes, including clothes and hair, from performer images using implicit distribution fields. Fig. 2 illustrates an overview of our method. In our work, the clothed human mesh is represented by smooth occupancy (Sec. 3.3). First, we elaborate on our observations about uncertainty (Sec. 3.2). In particular, a simple yet effective way is proposed to use a distribution-guided implicit network to learn the implicit distribution of each query point (Sec. 3.1), named D-IF. To train D-IF, we further propose an uncertainty distribution loss to constrain the predicted distribution (Sec. 3.4). Detailed information on the key designs of our approach is presented in the following sections.

3.1. Learning the implicit distribution filed

The aim of this work is to reconstruct a 3D human mesh with details on clothes from a single image. Further than predicting the point-wise value of implicit fields like the occupancy field [28] and the signed distance field (SDF) [32], we tend to infer the probability distribution of the implicit value for every point by a neural network, which will be

proven to maintain appropriate uncertainty but improve the reconstruction accuracy. The point-wise sampling of the predicted implicit distribution fields will regress a classical implicit field to represent the target surface.

We extract the same local deep features as ICON [43], and use them to predict the implicit distribution of each queried point. The extracted features include SMPL-body surface normal $N_{\text{body}} \in \mathbb{R}^3$, clothed surface normal $N_{\text{clothed}} \in \mathbb{R}^3$, and the signed distance value to the SMPL body $\text{SDF}_{\text{body}} \in \mathbb{R}$. The final 7D pixel-aligned feature $F_{7\text{D}}(p)$ is constructed by concatenating ($\oplus$) $N_{\text{body}}, N_{\text{clothed}}$ and SDF_{body}:

$$F_{7\text{D}}(p) := N_{\text{body}} \oplus N_{\text{clothed}} \oplus \text{SDF}_{\text{body}}, \tag{1}$$

While taking $F_{7\text{D}}(p)$ as inputs to predict directly the deterministic occupancy value is a viable option, as previously discussed, uncertainty should also be considered. To this end, we present a framework D-IF involving learnable implicit distribution fields to predict point-wise occupancy values with uncertainty. Specifically, for each query point p, an MLP designed as the Distribution Predictor in D-IF is trained to learn the distribution $P_\varphi(F_{7\text{D}}(p))$ which is assumed to be a Gaussian distribution [7]:

$$\begin{aligned} O_s(p) &\sim P_\varphi(F_{7\text{D}}(p)) = \mathcal{N}\left(\mu_\varphi(p), \sigma_\varphi(p)\right) \\ f(O(p) = Y) &= \frac{1}{\sigma_\varphi(p)\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{Y-\mu_\varphi(p)}{\sigma_\varphi(p)}\right)^2}. \end{aligned} \tag{2}$$

where probability density function f of the occupancy value formulates the conventional Gaussian distribution $\mu_\varphi(p)$ and $\sigma_\varphi(p)$ are the mean and the variance of the predicted distribution at the query point p. Through the learning of point-wise distribution, the coarse-level occupancy $O_s(p)$ will be sampled from the learned distribution.

Coarse-level occupancy from sampling maintains the uncertainty but lacks the accuracy to reconstruct the ground-truth mesh. To keep the balance between uncertainty and accuracy, D-IF designs an additional MLP as the Occupancy Rectifier R_δ to modify the coarsely sampled value into a fine result. The Occupancy Rectifier will concatenate ($\oplus$) the 7D features $F_{7\text{D}}$, the mean $\mu_\varphi(p)$ and variances $\sigma_\varphi(p)$ predicted from $P_\varphi(F_{7\text{D}}(p))$, and the sampled coarse occupancy $O_s(p)$ as input. All these features will be used to estimate the residual $\Delta O_s(p)$, which will be added onto the coarse occupancy $O_s(p)$, and get the fine occupancy results $\widetilde{O_s}(p)$ by the small amendments,

$$\begin{aligned} \Delta O_s(p) &= R_\delta\left(O_s(p) \oplus P_\varphi(F_{7\text{D}}(p)) \oplus F_{7\text{D}}(p)\right), \\ \widetilde{O_s}(p) &= O_s(p) + \Delta O_s(p), \end{aligned} \tag{3}$$

where Occupancy Rectifier is a MLP parameterized by δ.

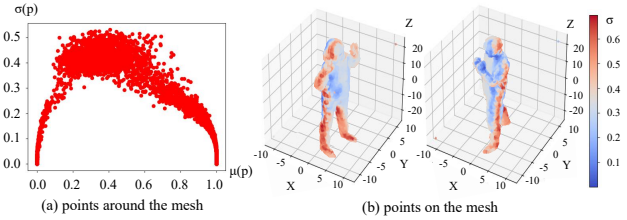

Figure 3. **Spatial-aware Uncertainty.** For a) points around the mesh and b) points on the surface. The degree of uncertainty increases as points approach the surface, and the body sides, which are typically with higher articulation than the torso.

3.2. Discussion on spatial-aware uncertainty

In line with traditional Bayesian deep learning principles [10, 29], we adopt a two-fold perspective on uncertainty during the prediction of the implicit field. Firstly, there exists *epistemic (model) uncertainty* [12] stemming from the idealized network architecture design and incomplete supervision through blend loss functions. This form of uncertainty signifies an inherent limitation which is not specifically addressed in our method. However, our primary focus lies on *aleatoric (data) uncertainty* derived from the data itself [17], reflecting random turbulence in geometric details present on a clothed human shape. Rather than homoscedastic uncertainty with constant strength irrespective of data properties, we emphasize heteroscedastic uncertainty that varies based on input data. Moreover, we contend that the uncertainty associated with the implicit value — specifically, the variance within the predicted distribution — is influenced by the spatial relationship between a query point and the potentially reconstructed surface.

Our experiments support this assumption. Specifically, we incorporate the data-dependent Bayesian Loss [17] to facilitate the learning of heteroscedastic uncertainty:

$$\mathcal{L}_B = \frac{1}{2N}\sum_{i=1}^{N} \frac{1}{\sigma(p_i)^2} \left\| \hat{O}(p_i) - O(p_i) \right\|^2 + \log \sigma(p_i), \tag{4}$$

where the first term measures the Mahalanobis distance [27] between the ground truth value and the value sampled from the predicted distribution, while the last regularization term of variance σ prevents the uncertainty from exploding. When the D-IF was only trained with the data-dependent Bayesian Loss, the variance at query points displays a declining tendency as the distance from the point to the target surface increases. Please find the illustration and explanation of the "spatial-aware uncertainty" from Fig. 3

Based on the above assumption and observations, we propose a loss design to model the uncertainty, termed the **uncertainty distribution loss**, which enables effective supervision of the variance in implicit distribution fields.

	Methods	Smooth Occupancy	Uncertainty Dist. Loss	CAPE-FP			CAPE-NFP			CAPE		
				Chamfer ↓	P2S ↓	Normals ↓	Chamfer ↓	P2S ↓	Normals ↓	Chamfer ↓	P2S ↓	Normals ↓
	Ours	✓	✓	0.684	0.677	0.048	0.838	0.821	0.055	0.785	0.771	0.050
A	PIFu* [37]	×	×	2.525	1.905	0.155	4.143	2.773	0.202	3.603	2.484	0.186
A	PaMIR* [47]	×	×	1.517	1.331	0.098	1.768	1.450	0.102	1.684	1.410	0.101
A	ICON [43]	×	×	0.775	0.715	0.054	1.004	0.930	0.063	0.928	0.859	0.060
A	ECON [42]	×	×	0.912	0.907	0.037	0.926	0.917	0.037	0.921	0.914	0.037
B	$\text{Ours}_{\text{D-IF (w/o Rectifier)}}$	×	×	0.976	0.900	0.064	1.245	1.124	0.075	1.155	1.049	0.071
B	$\text{Ours}_{\text{D-IF}}$	×	×	0.721	0.691	0.050	0.921	0.880	0.057	0.854	0.817	0.055
B	$\text{Ours}_{\text{D-IF}}$	✓	×	0.712	0.698	0.051	0.900	0.870	0.060	0.838	0.813	0.057
C	Ours_{L2}	✓	L2 for μ	0.833	0.777	0.065	1.019	0.947	0.070	0.957	0.890	0.068
C	Ours_{KL}	✓	Constant σ	0.708	0.689	0.051	0.885	0.857	0.057	0.826	0.801	0.055

Table 1. **Quantitative evaluation.** (A) performance w.r.t. SOTA; and the ablation studies of (B) Implicit distribution fields (D-IF), Smooth Occupancy, and (C) Uncertainty distribution loss. Notably, the best two results are colored as first second .

Method	CAPE-FP			CAPE-NFP			CAPE		
	Chamfer ↓	P2S ↓	Normals ↓	Chamfer ↓	P2S ↓	Normals ↓	Chamfer ↓	P2S ↓	Normals ↓
PIFu* [37]	2.525	1.905	0.155	4.143	2.773	0.202	3.603	2.484	0.186
PIFu_{dis}	1.867 (**-26%**)	1.263 (**-34%**)	0.111 (**-28%**)	3.413 (**-18%**)	1.985 (**-28%**)	0.169 (**-16%**)	2.898 (**-27%**)	1.744 (**-30%**)	0.150 (**-19%**)
PaMIR* [47]	1.517	1.331	0.098	1.768	1.450	0.102	1.684	1.410	0.101
$\text{PaMIR}_{\text{dis}}$	1.421 (-6%)	1.175 (-12%)	0.093 (-5%)	1.612 (-9%)	1.300 (-10%)	0.092 (-10%)	1.548 (-8%)	1.258 (-11%)	0.092 (-9%)
ICON [43]	0.775	0.715	0.054	1.004	0.930	0.063	0.928	0.859	0.060
ICON_{dis}	0.723 (-7%)	0.713 (-0.2%)	0.052 (-4%)	0.900 (-10%)	0.877 (-6%)	0.060 (-5%)	0.841 (-9%)	0.822 (-4%)	0.058 (-3%)

Table 2. **Generalizability proof on CAPE.** As a "plug-and-play" module, we apply implicit distribution fields (w/o smooth occupancy or uncertainty loss) to other methods, denoted as "X_{dis}". We denote the "simulated" SOTA methods with *.

3.3. Representing mesh with smooth occupancy

Following the above discussions, we tend to supervise the distribution of implicit values, like the occupancy field, by considering the relation between the relative location of query points and the associated uncertainty. To achieve this objective, an appropriate data structure informing the relative location of query points is imperatively. Conventional occupancy is typically binary, depicting whether a voxel is occupied by the interior of a surface or not. The binarity forgets all the quantitative information of the relative locations, which limits the spatial cost of computations but sacrifices the accuracy, especially for complicated shapes with fine details. While taking an accurate SDF as the representation would like to be computationally expensive when dealing with large-scale or high-resolution datasets. Moreover, neither of them is differentiated at points on the ground-truth surface. Therefore, in this paper, we adopt the smooth occupancy field to express a surface implicitly, which is also defined as the density field [30] in some contexts.

In particular, for a closed orientable surface $M \subset \mathbb{R}^3$ and a query point $p \in \mathbb{R}^3$, the spatial-aware smooth occupancy value $O(p|M)$ of M at p is defined as:

$$O_{gt}(p) = O_\alpha(p|M) := \frac{1}{1 + \exp(-\alpha \cdot \text{SDF}(p|M))}, \quad (5)$$

where $\text{SDF}(p|M)$ is the signed distance from p to M and α is the hyper-coefficient controlling the graininess of the voxel, where O_∞ degenerates into the classical discrete (binary) occupancy field when $\alpha \to \infty$. It is worth mentioning the coefficient α can be either determined by priors or settled as the learnable parameter for training.

Smooth occupancy field provides a smooth mapping $O_\alpha(\cdot|M) : \mathbb{R}^3 \mapsto [0, 1]$ with the level set $O_\alpha^{-1}(0.5) = M$, which endows the implicit value of points near the surface M with smoother gradients. In contrast to vanilla occupancy fields and signed distance functions (SDF), the smooth occupancy field exhibits differentiability and provides enhanced distance information. Finally, we apply Marching Cube [24] on the learned smooth occupancy field at the level set of $O_\alpha^{-1}(0.5)$, to extract the triangle mesh.

3.4. Uncertainty distribution loss

Training the D-IF network is nontrivial due to the necessity of effectively learning about uncertainty. With the smooth occupancy field, we can establish a connection between the variance of the distribution and spatial information through the uncertainty distribution loss.

Merely employing L1 loss to supervise the predicted occupancy distribution $P_\varphi(p)$ with the ground-truth smooth occupancy $O_{gt}(p)$ can yield accurate reconstructions by enforcing a sharp peak around the mean $\mu_\varphi(p)$. However, neglecting the supervision of variances may cause the network to converge to the Dirac impulse function [11]. Therefore, we incorporate uncertainty into the learning process in our design, which encourages the network to learn randomness. Meanwhile, we attempt to involve a relatively higher uncertainty in the coarse prediction, which exploits the capacity of the Occupancy Rectifier to regress a finer result.

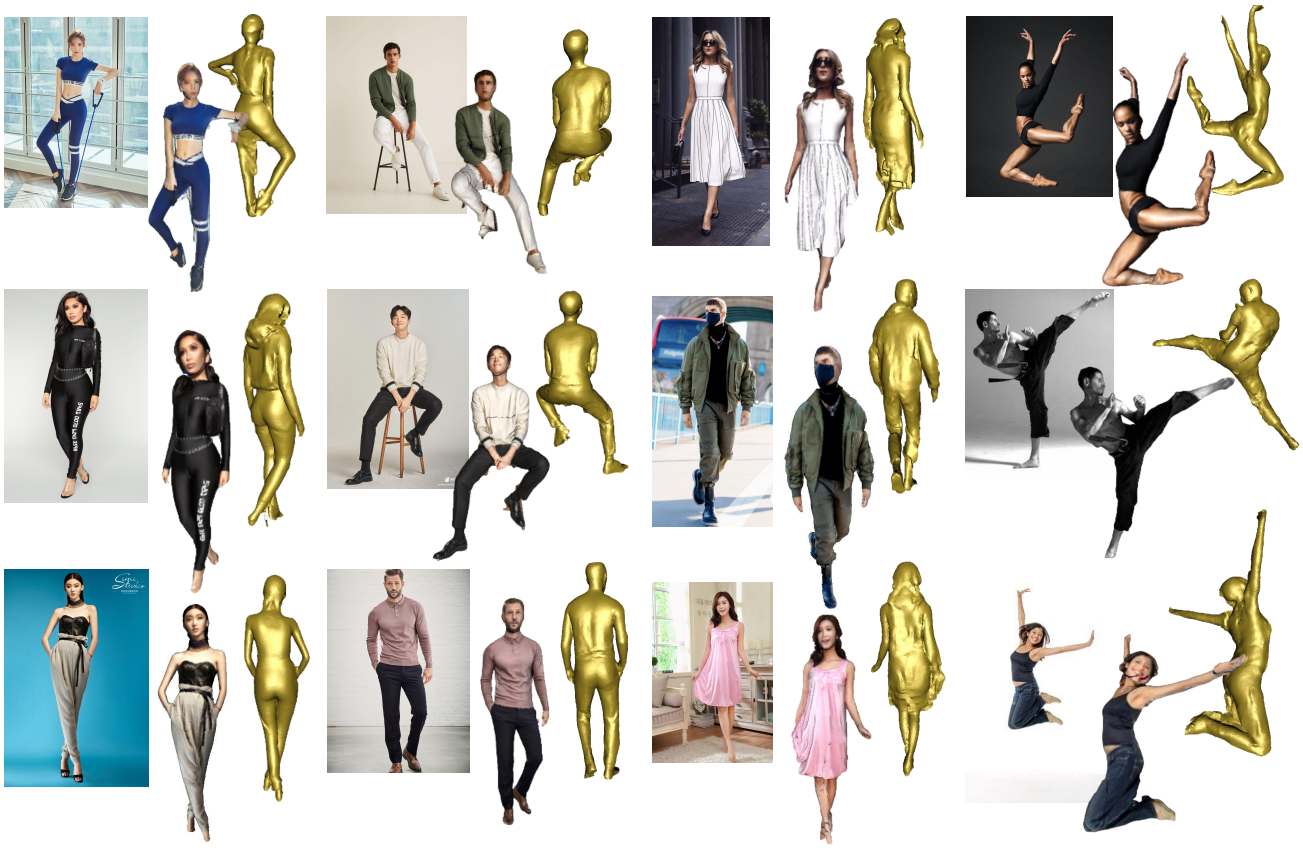

Figure 4. D-IF results for in-the-wild images with various clothing and challenging poses. **Zoom in** to see the geometric details.

To address this, we design an uncertainty distribution loss $\mathcal{L}_{\text{dis}}$ to balance appropriate uncertainty and more accurate sampling values where we carefully consider their interplay. Specifically, we design an expected distribution $\hat{P}(p) = \mathcal{N}(\mu_d(p), \sigma_d(p))$ with the a mean and variance μ_d and σ_d by taking the point-to-mesh distance into account. The mean of the designed distribution $\mu_d(p)$ for the query point is defined naturally as the ground-truth smooth occupancy $O_{gt}(p)$. As illustrated in Fig. 2, $\sigma_d(p)$ is designed to be negatively correlated with the point-to-mesh distance by

$$\sigma_d(p) = ke^{-\beta(\mu-0.5)^2}, \tag{6}$$

where k and β are hyperparameters to control the behavior of the uncertainty.

Consequently, we adopt the KL-divergence [22] to measure the distribution disparity from the perdition to the designed ground truth,

$$\mathcal{L}_{\text{dis}} = \text{KL}(P_\varphi(p) \,||\, \hat{P}(p)). \tag{7}$$

The total uncertainty loss of our network $\mathcal{L}_{\text{un}}$ is the combination of the distribution and 3D reconstruction losses:

$$\begin{aligned} \mathcal{L}_{\text{rec}} &= || \widetilde{O_s}(p) - O_{gt}(p) ||^2, \\ \mathcal{L}_{\text{un}} &= \alpha_1 \mathcal{L}_{\text{dis}} + \alpha_2 \mathcal{L}_{\text{rec}}. \end{aligned} \tag{8}$$

where α_1, α_2 are fixed loss weighting factors.

4. Experiment

Training. All the baselines are trained on THuman2.0 [46], which contains 525 high-quality human-textured scans in various poses captured by a dense DSLR rig along with their corresponding SMPL-X fits. The model is firstly trained with $\mathcal{L}_{\text{rec}}$ for 10 epochs, the learning rate is 1×10^{-4}. Then, we replace $\mathcal{L}_{\text{rec}}$ with uncertainty loss $\mathcal{L}_{\text{un}}$, and fine-tune the model for 5 epochs with the same learning rate. Empirically, we set $\alpha_1 = 1.0$, $\alpha_2 = 0.55$, $k = 0.6$ and $\beta = 4$. A larger q leads to a higher level of accuracy in the reconstruction process. However, an extremely large q will degenerate the smooth occupancy into a discrete 0-1 occupancy, which can harm the quality due to the loss of distance information. So we take a balance and set $q = 1 \times 10^3$. The entire training process takes approximately 1 day on a single NVIDIA RTX 3090 GPU with 36.5M learning parameters.

Testing. We mainly follow the test setting of ICON [43], which totally selects 150 scans from CAPE [26] to evaluate the reconstruction accuracy under challenging poses ("CAPE-NFP", 100 scans), and fashion poses ("CAPE-FP", 50 scans). The testing RGB images are rendered by rotating a virtual camera around the textured scans by $\{0°, 90°, 180°, 270°\}$. During the evaluation, we randomly sample the coarse occupancy value from the predicted distribution into Occupancy Rectifier to reconstruct the final prediction.

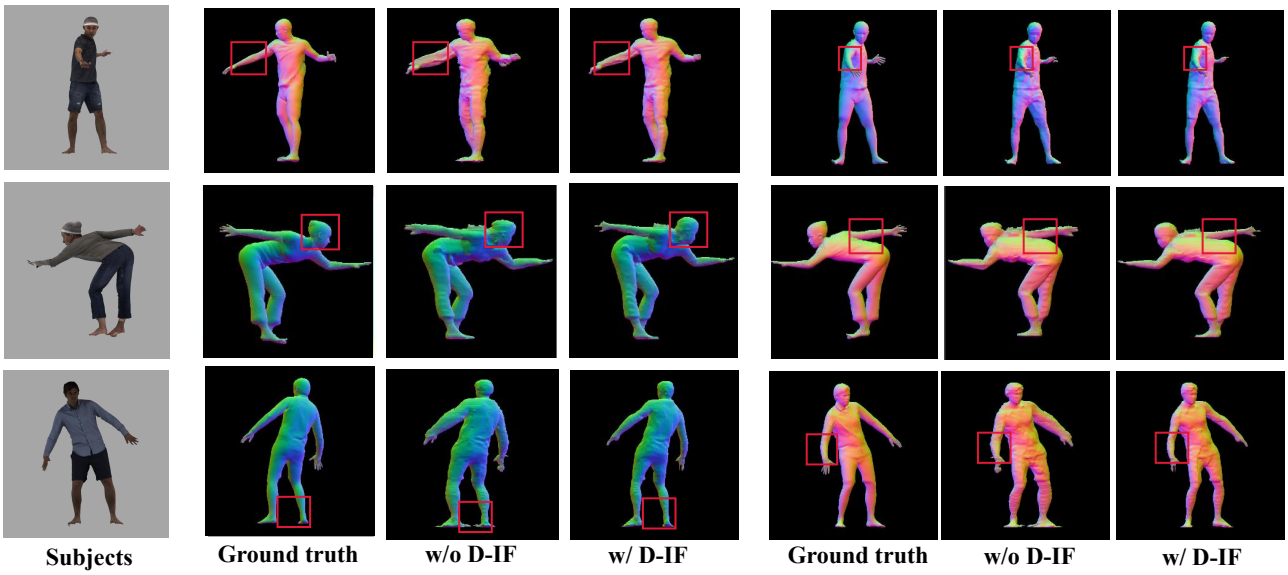

Figure 5. Reconstructions w/ and w/o implicit distribution field (D-IF). **Zoom in** to see the geometric details.

Metrics. "Chamfer" and "P2S" mainly capture the coarse geometric errors, while "Normals" mainly measure the high-frequency difference:

- **P2S distance (cm)**. The point-to-surface (P2S) distance is computed from randomly sampled scan points to their closest face on the reconstructed mesh.
- **Chamfer distance (cm)**. This metric could be regarded as *bilateral* P2S distance, which also considers the P2S distance from randomly sampled points on the reconstructed mesh to the ground-truth scans.
- **Normals difference (L2)**. The normal images used for evaluation are rendered from the reconstructed and ground-truth meshes, by rotating a virtual camera around them by $\{ 0°, 90°, 180°, 270° \}$.

4.1. Comparison with state-of-the-art methods.

We conduct quantitative comparisons with mainstream SOTAs, including PIFu [37], PaMIR [47], ICON [43] and ECON [42]. For a fair comparison, we follow the re-implementations of PIFu and PaMIR in ICON [43], and re-train all models on THuman2.0 [46].

The quantitative comparison results are presented in Tab. 1-A. Our proposed method has demonstrated superior performance compared to all baselines, confirming its superiority by a significant margin. Notably, our method has achieved even better results than the previous best-performing method, ECON [42], on the challenging "CAPE-NFP" dataset. This dataset evaluates the model's ability to reconstruct clothed humans with out-of-distribution poses. The impressive performance of our method on this challenging dataset highlights its robustness and effectiveness in handling complex real-world scenarios.

The results in Fig. 1 clearly demonstrate that our approach produces more accurate and detailed reconstructions. This aligns with our objective of using uncertainty distribution instead of deterministic occupancy values to represent the geometries. The superiority of our method can be attributed to the innovative point-wise distribution learning, supplemented by the inclusion of uncertainty loss.

In summary, these two findings discussed above provide strong evidence that uncertainties exist in the clothed human reconstruction process and that they follow explainable laws. These observations confirm the validity of our motivation to leverage point-wise distribution learning to better capture uncertainty. By providing a more accurate representation of uncertainty in the reconstruction process, our approach has the potential to improve the quality and reliability of clothed human mesh reconstruction. Overall, our results support the use of uncertainty learning in this context and highlight the importance of considering uncertainty in machine learning applications. More reconstruction results for in-the-wild images are shown in Fig. 4.

4.2. Ablation Study.

We further validate the effectiveness of 1) implicit distribution field (D-IF) in Tab. 1-B, 2) Occupancy Rectifier in Tab. 3, and 3) uncertainty distribution loss in Tab. 1-C.

Implicit Distribution Field (D-IF). As shown in Tab. 1-B, we replace "Value-based Implicit Function (IF)" used by competitors with our proposed "Implicit Distribution Field (D-IF)", and benchmark their performance under original

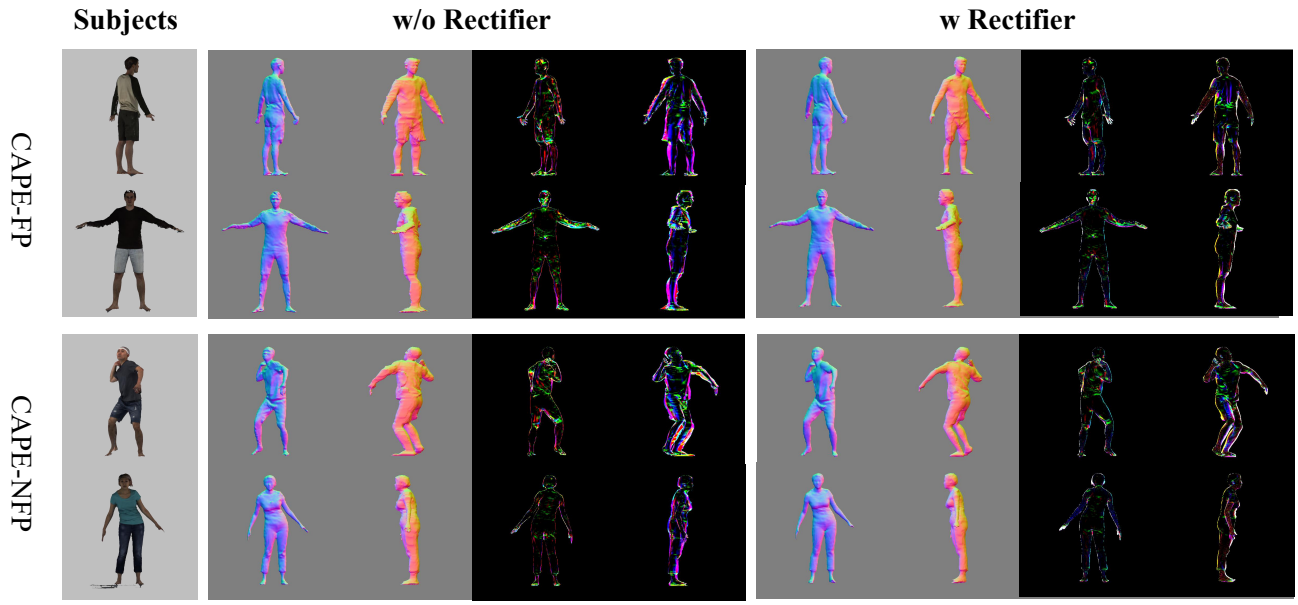

Figure 6. Error map for the reconstructed mesh w/ and w/o Occupancy Rectifier. **Zoom in** to see the geometric details.

occupancy / smooth occupancy. We find that uncertainty distribution has significantly improved the reconstruction quality, especially for stretching gestures, and also reduce the abnormal artifacts of limbs; see Fig. 5.

	Rectifier	P2S (cloth) ↓	Chamfer (cloth) ↓	P2S (body) ↓	Chamfer (body) ↓
Loose	×	1.033	1.090	1.268	1.377
	✓	0.955 (-7%)	0.927 (-14%)	0.843 (-33%)	0.864 (-37%)
Tight	×	0.910	0.988	0.939	1.035
	✓	0.586 (-35%)	0.607 (-38%)	0.639 (-31%)	0.669 (-35%)

Table 3. **Ablation experiments (w/ Rectifier vs. w/o Rectifier) on both loose and tight clothing.** Metrics (cloth) are calculated between the reconstructed surface and ground truth *clothed* surface, and metrics (body) are calculated between the reconstructed surface and ground truth *body* surface. The results indicate that the Occupancy Rectifier enhances reconstruction accuracy mainly by bringing the predicted mesh closer to the SMPL body mesh.

Occupancy Rectifier. Besides, we also demonstrate the necessity of Occupancy Rectifier in Tab. 1-B. Results show that removing the Rectifier ($\text{Ours}_{\text{D-IF (w/o Rectifier)}}$) finally harms the reconstruction quality, increasing Chamfer error from 0.854 cm to 1.155 cm (35% increase), and see normal error maps (w/ Rectifier vs. w/o Rectifier) in Fig. 6.

To further validate the impact of the Occupancy Rectifier, we conducted ablation studies (w/ vs w/o Rectifier) on both loose and tight clothing categories in the CAPE dataset. The categorization was based on the "P2S" and "chamfer" distances (measured in centimeters) between the reconstructed surface and ground truth clothed mesh or minimally-clothed mesh. The evaluation results presented in Tab. 3 demonstrate that, for subjects wearing loose clothing, there were significant reductions in P2S (body, 33% decrease) and Chamfer (body, 37% decrease) when employing the Occupancy Rectifier. However, there were comparatively smaller decreases observed in P2S (cloth 7% decrease) and Chamfer (cloth, 14% decrease). For subjects wearing tight clothing, both Chamfer (body, 35% decrease) and Chamfer (cloth, 38% decrease) exhibited substantial improvements. These findings suggest that the Rectifier primarily enhances reconstruction accuracy by bringing the predicted mesh closer to the SMPL body mesh.

Uncertainty Distribution Loss. Finally, as shown in Tab. 1-C, we compare the uncertainty distribution loss to other versions, which simply regress the predicted μ with L2 loss or change the variance of the target distribution into a constant (flat distribution). The chosen constant is the surface value of the designed distribution, in Eq. (6). The results show that the introduction of the designed distribution in uncertainty loss does improve the robustness of reconstruction, as described in Sec. 3.4.

4.3. Generalization of D-IF.

The core module, Implicit Distribution Field (D-IF), can be easily integrated into other implicit-based approaches, like PIFu [37], PaMIR [47] and ICON [43]. We re-train the above three baselines with our D-IF, see Tab. 2 for the quantitative comparison. D-IF boosts all the baselines on all metrics, especially for PIFu, where D-IF even reduces the reconstruction error by 27% in Chamfer, 30% in P2S, and 19% in Normals. Additionally, by incorporating uncertainty, D-IF improves the recovery of non-rigid deformations in clothing, as shown in non-fashion poses (NFP) in Tab. 2. All these results have demonstrated that D-IF generalizes quite well to a wide range of implicit-based human reconstruction approaches. Hence, we have reason to believe that D-IF, as a powerful *plug-and-play* module, could also be used to represent other non-human shapes.

5. Conclusion

Clothed human reconstruction is a fundamental task in human digitization. In this paper, we propose a novel method to learn the uncertainty that exists in clothed human reconstruction process for better geometric details. In our work, a new implicit representation named smooth occupancy field is introduced to represent clothed humans in neural space. Then, a distribution-guided implicit function network is proposed to learn point-wise distribution to recover the occupancy distribution field. Finally, a novel uncertainty loss is presented to better train the network. Experimental results demonstrate that our method achieves state-of-the-art performance and can be generalized to other baselines to consistently improve their performances.

Acknowledgments. Xueting Yang, Zhaoxin Fan, Wei Wang and Hao Xu are supported by Psyche AI Inc. Yihao Luo is partly funded by the Imperial College London and partly supported by Psyche AI Inc. Yuliang Xiu is funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No.860768 (CLIPE).

References

[1] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 3

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *International Conference on 3D Vision (3DV)*, 2018.

[3] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 3

[4] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic Monocular 3D Reconstruction of Humans Wearing Clothing. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 3

[5] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Multi-view depth estimation by fusing single-view depth probability with multi-view geometry. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[6] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *International Conference on Computer Vision (ICCV)*, 2019. 3

[7] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Springer, 2006. 4

[8] Wentao Chao, Xuechun Wang, Yingqian Wang, Liang Chang, and Fuqing Duan. Learning Sub-Pixel Disparity Distribution for Light Field Depth Estimation. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[9] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[10] John Denker and Yann LeCun. Transforming neural-net output levels to probability distributions. *Conference on Neural Information Processing Systems (NeurIPS)*, 3, 1990. 4

[11] Paul Adrien Maurice Dirac. The quantum theory of the electron. *Proceedings of the Royal Society of London.*, 1928. 5

[12] Alex Graves. Practical variational inference for neural networks. *Conference on Neural Information Processing Systems (NeurIPS)*, 24, 2011. 4

[13] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 1, 3

[14] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. ARCH++: Animation-Ready Clothed Human Reconstruction Revisited. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[15] Yang Hong, Juyong Zhang, Boyi Jiang, Yudong Guo, Ligang Liu, and Hujun Bao. Stereopifu: Depth aware clothed human digitization via stereo vision. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[16] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[17] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Conference on Neural Information Processing Systems (NeurIPS)*, 30, 2017. 4

[18] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-degree textures of people in clothing from a single image. In *International Conference on 3D Vision (3DV)*, 2019. 1, 3

[19] Ruilong Li, Kyle Olszewski, Yuliang Xiu, Shunsuke Saito, Zeng Huang, and Hao Li. Volumetric human teleportation. In *ACM SIGGRAPH 2020 Real-Time Live*, 2020. 1

[20] Ruilong Li, Yuliang Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li. Monocular real-time volumetric performance capture. In *European Conference on Computer Vision (ECCV)*, 2020. 1

[21] Tingting Liao, Xiaomei Zhang, Yuliang Xiu, Hongwei Yi, Xudong Liu, Guo-Jun Qi, Yong Zhang, Xuan Wang, Xiangyu Zhu, and Zhen Lei. High-Fidelity Clothed Avatar Reconstruction from a Single Image. In *Computer Vision and Pattern Recognition (CVPR)*, June 2023. 1

[22] Friedrich Liese and Igor Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 2006. 2, 6

[23] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *Transactions on Graphics (TOG)*, 2015. 1, 3

[24] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 1987. 2, 5

[25] Qianli Ma, Shunsuke Saito, Jinlong Yang, Siyu Tang, and Michael J Black. Scale: Modeling clothed humans with a surface codec of articulated local elements. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[26] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J Black. Learning to dress 3d people in generative clothing. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 3, 6

[27] Prasanta Chandra Mahalanobis. On the generalized distance in statistics. *Sankhy: The Indian Journal of Statistics, Series A (2008)*, 2018. 4

[28] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3

[29] RM Neal. Bayesian learning for neural networks [Ph.D. thesis]. *Toronto, Ontario, Canada: Department of Computer Science, University of Toronto*, 1995. 4

[30] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 5

[31] Liang Pan, Xinyi Chen, Zhongang Cai, Junzhe Zhang, Haiyu Zhao, Shuai Yi, and Ziwei Liu. Variational relational point completion network. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[32] Jeong Joon Park, Peter Florence, Julian Straub, Richard

Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3

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

[34] Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3d object detection. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[35] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[36] Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[37] 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 (ICCV)*, 2019. 1, 3, 5, 7, 8

[38] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 3

[39] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Photo wake-up: 3d character animation from a single photo. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[40] Donglai Xiang, Timur Bagautdinov, Tuur Stuyck, Fabian Prada, Javier Romero, Weipeng Xu, Shunsuke Saito, Jingfan Guo, Breannan Smith, Takaaki Shiratori, Yaser Sheikh, Jessica Hodgins, and Chenglei Wu. Dressing avatars: Deep photorealistic appearance for physically simulated clothing. *Transactions on Graphics (TOG)*, 41(6), nov 2022. 1

[41] Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica Hodgins. Monoclothcap: Towards temporally coherent clothing capture from monocular rgb video. In *International Conference on 3D Vision (3DV)*, 2020. 1

[42] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Obtained from Normals. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 3, 5, 7

[43] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed Humans Obtained from Normals. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[44] Ze Yang, Shenlong Wang, Siva Manivasagam, Zeng Huang, Wei-Chiu Ma, Xinchen Yan, Ersin Yumer, and Raquel Urtasun. S3: Neural shape, skeleton, and skinning fields for 3d human modeling. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[45] Ruolin Ye, Wenqiang Xu, Haoyuan Fu, Rajat Kumar Jenamani, Vy Nguyen, Cewu Lu, Katherine Dimitropoulou, and Tapomayukh Bhattacharjee. Rcareworld: A human-centric simulation world for caregiving robots. *International Conference on Intelligent Robots and Systems (IROS)*, 2022. 1

[46] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 6, 7

[47] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. PaMIR: Parametric Model-Conditioned Implicit Representation for Image-based Human Reconstruction. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 1, 3, 5, 7, 8

[48] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 3