FOF-X: Towards Real-time Detailed Human Reconstruction from a Single Image

Qiao Feng[†], *Student Member, IEEE,* Yuanwang Yang[†], *Student Member, IEEE,* Yebin Liu, *Member, IEEE,* Yu-Kun Lai, *Senior Member, IEEE,* Jingyu Yang, *Senior Member, IEEE,* and Kun Li[*], *Senior Member, IEEE*

Abstract—We introduce FOF-X for real-time reconstruction of detailed human geometry from a single image. Balancing real-time speed against high-quality results is a persistent challenge, mainly due to the high computational demands of existing 3D representations. To address this, we propose Fourier Occupancy Field (FOF), an efficient 3D representation by learning the Fourier series. The core of FOF is to factorize a 3D occupancy field into a 2D vector field, retaining topology and spatial relationships within the 3D domain while facilitating compatibility with 2D convolutional neural networks. Such a representation bridges the gap between 3D and 2D domains, enabling the integration of human parametric models as priors and enhancing the reconstruction robustness. Based on FOF, we design a new reconstruction framework, FOF-X, to avoid the performance degradation caused by texture and lighting. This enables our real-time reconstruction system to better handle the domain gap between training images and real images. Additionally, in FOF-X, we enhance the inter-conversion algorithms between FOF and mesh representations with a Laplacian constraint and an automaton-based discontinuity matcher, improving both quality and robustness. We validate the strengths of our approach on different datasets and real-captured data, where FOF-X achieves new state-of-the-art results. The code has already been released for research purposes at https://cic.tju.edu.cn/faculty/likun/projects/FOFX/index.html.

Index Terms—real-time, 3D human reconstruction, single image, monocular 3D reconstruction.
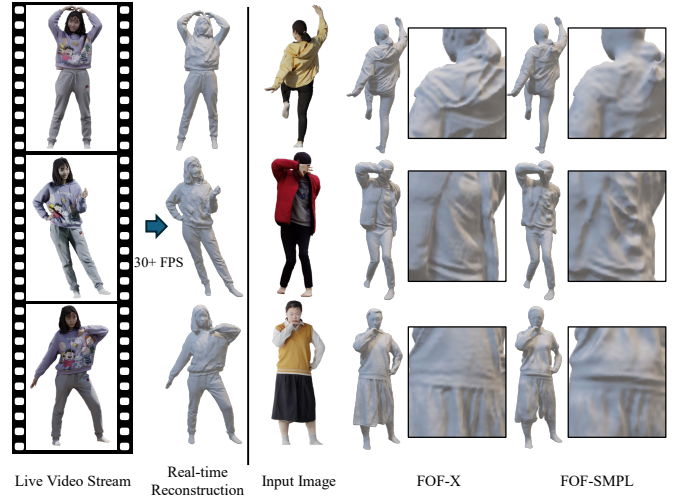
Fig. 1. Left: Our FOF-X can reconstruct 3D human shapes from a live video stream with a real-time speed of over 30 FPS. Right: Compared to the original FOF-SMPL, our FOF-X demonstrates better robustness to variations in texture and lighting. Under challenging lighting conditions, such as strong illumination or shadows, FOF-X produces more detailed and accurate reconstructions (first two rows). FOF-X effectively avoids the incorrect reconstruction caused by textures, such as the stripe pattern on the edge of the shirt (third row). Note that FOF-SMPL is not real-time.

I. INTRODUCTION

RECONSTRUCTING a 3D human from a single image has emerged as a popular task in computer vision and graphics, which can be widely used in various downstream applications, such as mixed reality and virtual try-on. However, real-time, high-fidelity monocular 3D human reconstruction remains challenging. The core of this challenge lies in the 3D representation, as it significantly influences the design and performance of reconstruction and generation approaches. Despite some promising results, existing methods [1]–[4] typically suffer from high computational consumption and lack of robustness. Consequently, a good 3D geometry representation is essential for 3DTV, Holographic Telepresence

† Equal contribution.
* Corresponding author: Kun Li (E-mail: lik@tju.edu.cn).
Qiao Feng, Yuanwang Yang and Kun Li are with the College of Intelligence and Computing, Tianjin University, Tianjin 300350, China. E-mail: {fengqiao,yyw,lik}@tju.edu.cn
Yebin Liu is with the Department of Automation, Tsinghua University, Beijing 100084, China. E-mail: liuyebin@mail.tsinghua.edu.cn
Yu-Kun Lai is with the School of Computer Science and Informatics, Cardiff University, Cardiff CF24 4AG, United Kingdom. E-mail: Yukun.Lai@cs.cardiff.ac.uk
Jingyu Yang is with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China. E-mail: yjy@tju.edu.cn

systems, and other real-time applications, which must fulfill the requirements of accuracy, efficiency, and compatibility [5].

Classic representations, such as voxel grids [6] or meshes [7]–[9], have been explored for monocular human reconstruction. However, voxel grids require a space complexity of $O(n^3)$ (where n is the resolution of grids in each dimension), and meshes struggle with topology changes or large deformations. Although some recent methods [4], [10] propose optimization-based pipelines combining normal maps, voxel grids, and meshes, they all suffer from efficiency issues.

Implicit neural representations have emerged and been widely used in monocular human reconstruction [1], [3], [11]–[13]. These methods treat 3D space as a continuous field, such as an occupancy or signed distance field, represented as $F(x, y, z) : \mathbb{R}^3 \to \mathbb{R}$. Instead of voxel grids, a neural network models the field, allowing results at any resolution. However, these methods infer values across numerous spatial grid points, which slows down inference and makes high-frame-rate or real-time reconstruction a significant challenge. Based on PIFu [1], Monoport [14] proposes an efficient sampling scheme to speed up inference, achieving 15 FPS only for mesh-free rendering. However, when generating a complete mesh, their

TABLE I
COMPARISON WITH EXISTING 3D REPRESENTATIONS

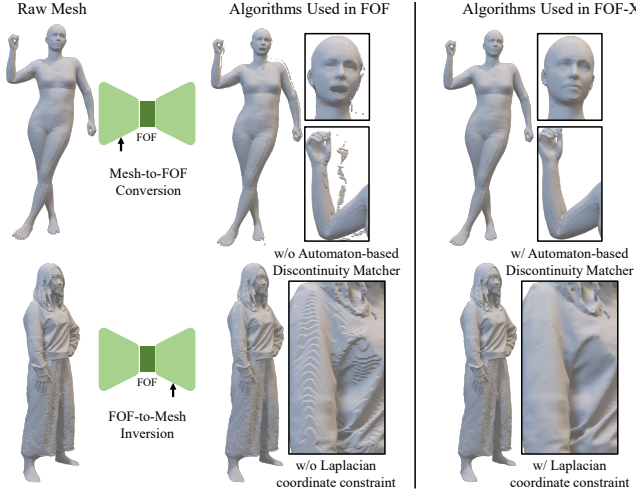| Representation	Aligned with Images	High-Quality	Computational Efficiency	Flexible Topology
Parametric Model [15]–[17]	✗	✗	✓	✗
Voxel Grid [6]	✓	✗	✗	✓
Mesh [7]–[9]	✗	✗	✓	✗
Implicit Function [1], [3], [11]–[13], [18]	✓	✓	✗	✓
Our FOF(-X)	✓	✓	✓	✓

Fig. 2. FOF and meshes can be inter-converted flexibly. The newly designed inter-conversion algorithms in FOF-X exhibit better robustness and quality. Our automaton-based discontinuity matcher eliminates floating artifacts during the conversion process (first row). With the Laplacian coordinate constraint, we resolve stair-step artifacts on the recovered meshes (second row).

method operates at less than 10 FPS, and the results are often worse than PIFu. These limitations hinder Monoport's applicability with downstream applications.

To address these challenges, we propose *Fourier Occupancy Field (FOF)*, a novel representation for monocular real-time and detailed 3D human reconstruction. FOF is an expressive, efficient, and flexible 3D geometric representation manifested as a 2D map aligned with the input image. The key idea of FOF is to represent a 3D object with a 2D field by decomposing the occupancy field along the z axis into a Fourier series, retaining only the first few terms. This approach is not only memory-efficient but also adept at preserving essential geometric information. By implementing grid sampling on the x and y axes, FOF can be explicitly stored as a multi-channel image. This feature allows 3D geometric priors, such as SMPL models, to be fed directly into the CNN. With our inter-conversion algorithms between FOF and mesh representations, 3D and 2D information can be processed within a unified framework. Unlike depth maps, FOF encapsulates the complete geometry of an object, not just its visible parts, significantly enhancing the fidelity of the reconstructed human shapes. Detailed comparisons with existing representations are presented in Table I.

When applying FOF directly to a real-time system, we observed certain limitations in robustness. The reconstruction is degraded by texture and lighting effects. Additionally, efficiency and robustness issues in the original mesh-to-FOF

conversion algorithm used in the conference version prevent the use of the SMPL prior in the real-time system, resulting in unsatisfactory performance on challenging poses. Furthermore, the original FOF-to-mesh algorithm comes with stair-step artifacts, further compromising the reconstruction quality. To address these challenges, we propose FOF-X, a novel reconstruction framework. With a single RGB image as input, FOF-X first translates it to dual-sided normal maps, which are unaffected by texture and lighting, as an internal representation of the human body. This strategy allows the network to focus on geometric details, significantly improving the robustness and performance of the reconstruction. To ensure the robustness and fidelity of FOF computations, we design an automaton-based discontinuity matcher to filter out invalid fragments. This matcher can be parallelized on a GPU, enabling real-time conversion of parametric models. Additionally, we incorporate a Laplacian coordinate constraint in the FOF-to-mesh conversion process to eliminate artifacts caused by view bias, resulting in more accurate mesh reconstructions and improving the fidelity of the final output. As shown in Fig. 2, the newly designed mesh-to-FOF and FOF-to-mesh algorithms in FOF-X effectively address the robustness issues. Based on these updates, our final method, FOF-X, achieves new state-of-the-art results in both speed and accuracy, contributing a novel solution for real-time detailed human reconstruction from a single image. Experiments also show that FOF-X has better generalization for real-captured data.

The main contributions of our work are as follows:

- We propose *Fourier Occupancy Field (FOF)*, a novel representation for 3D humans, which can represent a high-quality geometry with a 2D map aligned with the image, bridging the gap between 2D images and 3D geometries.
- We introduce FOF-X, a new reconstruction framework designed for detailed 3D human reconstruction from a single RGB image. FOF-X utilizes dual-sided normal maps as an internal representation, avoiding the effects of texture and lighting to focus on geometric accuracy.
- In FOF-X, we design parallelized inter-conversion algorithms between FOF and meshes with a Laplacian coordinate constraint and an automaton-based discontinuity matcher, further enhancing the robustness and reconstruction quality.
- Compared with state-of-the-art methods based on other representations, our approach can produce high-quality results in real-time. Our system is the first 30+FPS pipeline, achieving a twofold improvement in speed over Monoport [14], coupled with superior quality.

An early version of our method, including only FOF, was published as a conference paper [19]. Based on our FOF representation, several works for generation [20] and reconstruction [21]–[23] have been proposed. In this paper, we substantially extend the original version by developing **FOF-X**. Firstly, we design a new reconstruction framework that greatly mitigates the performance degradation caused by texture and lighting effects. Secondly, we propose a robust mesh-to-FOF conversion algorithm with an automaton-based

discontinuity matcher, enabling real-time execution and significantly improving the system's robustness when facing challenging poses. Thirdly, we propose a FOF-to-mesh algorithm with a Laplacian coordinate constraint for greater robustness and fidelity. Such a strategy effectively addresses artifacts caused by view-direction bias without any loss of geometric details. Fourth, we employ cosine series as the subspace approximation basis, retaining the Fourier-like approximation capacity while eliminating Gibbs phenomena through even-periodic boundary extension. Additionally, we evaluate FOF-X through comprehensive experiments including: testing on more datasets, comparisons with various state-of-the-art methods, and more detailed ablation studies, all confirming the method's effectiveness, efficiency and robustness.

FOF-X offers three key advantages: (1) FOF-X is sampling-scalable along all three spatial axes, allowing for resolution adaptation at inference without retraining, which supports diverse deployment scenarios with varying speed, quality, and memory constraints; (2) FOF-X represents 3D objects as multi-channel images, requiring only a simple tensor multiplication to reconstruct the 3D occupancy field; (3) FOF-X can be seamlessly inter-converted to and from the mesh representation, making it compatible with traditional graphics pipelines. Additionally, as a compact and computationally efficient 3D representation, FOF-X is suitable for real-time applications like holographic transportation.

II. RELATED WORK

Regarding geometric representations, we categorize mainstream monocular 3D human reconstruction approaches into three types: surface-based reconstruction, volume-based reconstruction, and "sandwich-like" reconstruction. Note that "sandwich-like" reconstruction can be considered a blend of surface-based and volume-based methods. Additionally, generative models can also be used directly or indirectly for single-image 3D human reconstruction. We review these methods in this section.

A. Surface-based Reconstruction

Surface-based approaches focus on inferring the interface between geometry and empty space, and can be broadly classified into three categories: parametric models, UV-map-based methods, and graph-based methods.

Parametric models, such as SMPL [15], SMPL-X [16], and STAR [17], are popular representations for naked 3D humans. These models are derived from large datasets of naked human shapes and use statistical methods to create analytical models, which can generate a human body mesh with just a few dozen parameters. Building on these models, many works [24], [25] estimate 3D human shapes from an RGB image by predicting the parameters of the parametric model. However, these methods are limited in that they cannot reconstruct clothing or hair.

UV-map-based methods attempt to estimate clothed humans by adjusting the vertices of the parametric models. Alldieck et al. [8] warp the input image to align with the UV map and calculate the displacements on the T-pose SMPL mesh.

However, their results are always in T-pose and do not align well with the input images. Zhu et al. [7], [26] employ a four-stage process constrained by joints, silhouettes, and shading, but their results are sometimes inconsistent with the images.

Graph-based representations are naturally compatible with triangle meshes. Li et al. [9] use graph neural networks to reconstruct topology-consistent geometries of clothed humans, but the recovered mesh tends to be smooth and lacks fine details. Representing 2D manifold surfaces of different 3D objects with a fixed structure is very challenging. Consequently, most methods heavily rely on the topology defined by SMPL and cannot produce detailed human geometries with complex topology. Habermann et al. [27] propose a deep learning approach for monocular dense human performance capture, but this method requires scanning the actor with a 3D scanner to create a template mesh. In a word, graph-based methods struggle to align image features accurately with the results, leading to unstable and over-smoothed outputs.

B. Volume-based Reconstruction

Volume-based methods estimate attributes such as occupancy and signed distance for each point in the 3D space, allowing them to represent 3D shapes with arbitrary topology. Voxel grids and implicit neural networks are the two main representations used in volume-based methods.

Zheng et al. [6] regress voxel grids of 3D humans using convolutional neural networks, but this approach requires intensive memory and results in models with limited resolution. Unlike voxel grids, implicit neural representations can depict detailed 3D shapes without resolution limitations. Saito et al. [1] use pixel-aligned functions to reconstruct a 3D human from a single RGB image and further improve 3D geometric details using normal maps [11]. Zheng et al. [12] use a voxelized SMPL mesh as a prior to enhance robustness. Xiu et al. [3] propose a method to correct the SMPL model estimated by other methods [24], [25], [28] based on the input image, and then regress the occupancy field from the normal map and the signed distance field of the corrected SMPL. While their results are robust to human poses, the geometric details often do not match the input images and can be noisy.

Further advancements like GTA [29], SiFU [2], and D-IF [30] explore implicit neural representations with a globally-correlated 3D-decoupling transformer and uncertainty-aware implicit field inference, respectively. However, they still struggle with computational efficiency. Although Li et al. [14] propose an efficient sampling scheme to speed up inference, they achieve 15 FPS only for mesh-free rendering, and when a complete mesh needs to be generated, their method operates at less than 10 FPS.

Recent advancements have seen the utilization of Neural Radiance Fields (NeRF) [31] in various reconstruction approaches, and subsequent methods [32]–[34] have been further adapted for human reconstruction. NeRF applies volume rendering to implicit volume-based representations, enabling direct supervision of 3D scenarios with 2D images. Consequently, SHERF [33] achieves the first generalizable Human NeRF from a single human image, while ELICIT [34] employs

a pre-trained CLIP model [35] for contextual understanding. However, NeRF-based approaches primarily focus on view synthesis and often lack detailed geometry generation from monocular images.

C. "Sandwich-like" Reconstruction

In recent works [4], [36], [37], 3D objects are modeled using front and back depth maps, resembling a sandwich structure that encloses the inner space. However, depth maps only represent part of the geometric surface and cannot fully reconstruct the object. Therefore, these methods require volume-based post-processing, such as IF-Nets [38] and Poisson surface reconstruction [39], to complete the result further. Any-Shot GIN [37] extends the sandwich-like scheme to novel classes of objects by predicting the depth maps of two sides from a single image and performing shape completion with IF-Nets. ECON [4] follows a similar approach and extends it to 3D humans. It estimates the normal maps of two sides and converts them into depth maps using an optimization method. The parametric human model prior is also employed in the shape completion stage to "inpaint" the missing geometry. However, this optimization process is time-consuming, and its instability may lead to incorrect geometric estimates.

D. Leveraging Generative Models

Reconstructing a 3D human body from a single image is inherently challenging due to incomplete observations, making it difficult to regress geometric information accurately. In this case, generative models can offer plausible completion for invisible regions. SiTH [18] generates the back-view image with an image-conditioned diffusion module and then predicts dual-sided normal maps for subsequent reconstruction. TeCH [10] recovers human meshes with a SMPL-X initialized hybrid Deep Marching Tetrahedra (DMTet) representation and Score Distillation Sampling (SDS) Loss [40], achieving visually impressive results but with noisy surfaces. Human-LRM [41] first uses an LRM [42] model to reconstruct the human body as a NeRF from a single image. It then applies a diffusion model to refine the multi-view images rendered from the NeRF, followed by another LRM reconstruction based on the refined images to produce the final result, but does not provide a substantial improvement in reconstructed geometry quality. PSHuman [43] first uses a multi-view image diffusion model to generate six views of global full-body images and local face images. Then, using the generated normal and color maps, SMPL-X is deformed and remeshed by differentiable rasterization. However, the heavy pipeline makes it inefficient for real-time use. These generative model-based methods are typically time-consuming, and the reconstructed result does not accurately match the input image.

To solve the problems with existing representations, we present Fourier Occupancy Field (FOF) in this paper. FOF is well-aligned with the input image and offers superior efficiency compared to existing representations. Leveraging the proposed efficient and flexible 3D geometry representation of FOF, we design an extensive reconstruction framework FOF-X and parallelized inter-conversion algorithms between FOF

and meshes. Based on FOF-X, we contribute a robust high-quality human reconstruction system operating at 30 FPS with a single RGB camera.

III. METHOD

Fig. 3 illustrates the overall pipeline of our method. Our work aims to reconstruct a high-fidelity 3D human model from a single RGB image in real time. In this section, we elaborate on the technical details of our approach. We first introduce our vanilla *FOF (Fourier Occupancy Field)*, an efficient and flexible representation for 3D geometry, in Sec. III-A. Then, we extend *FOF* to *FOF-X* with a new reconstruction framework with dual-sided normal maps as an internal representation (Sec. III-B), a parallelized mesh-to-FOF conversion algorithm with an automaton-based discontinuity matcher (Sec. III-C1), and a more robust FOF-to-mesh extraction algorithm with a Laplacian coordinate constraint (Sec. III-C2). These updates further improve the quality of the reconstructed meshes while keeping the overall pipeline running in real time.

A. Formulation of Fourier Occupancy Field

1) Vanilla FOF: An overview of our FOF representation is shown in Fig. 4. Without loss of generality, 3D objects to be reconstructed are normalized into a $[-1,1]^3$ cube. Denoted by $(x,y,z) \in [-1,1]^3$ the 3D coordinates, and by S the surface of a 3D object, the 3D object can be described by a 3D occupancy field $F : [-1,1]^3 \mapsto \{0, 0.5, 1\}$, where the occupancy function $F(x,y,z)$ is defined as

$$F(x,y,z) = \begin{cases} 1, & (x,y,z) \text{ is inside the object,} \\ 0.5, & (x,y,z) \in S, \\ 0, & (x,y,z) \text{ is outside the object.} \end{cases} \quad (1)$$

Such a 3D representation is highly redundant, because only a small subset, *i.e.*, the iso-surface of $F(x,y,z)$ with value of 0.5, is sufficient to represent the surface of the object. Direct inference of the 3D occupancy field F from a single image not only confronts with the curse of dimensionality, but also requires more computation and memory in handling high dimensional feature maps. Note that the occupancy field defined in the 3D cube can be regarded as the collection of a large number of 1D signals defined on the lines along the view direction. Without loss of generality, we assume that the view direction is the same as the z-axis of the 3D cube. The occupancy function, as a 1D signal $f(z) : [-1,1] \mapsto \{0, 0.5, 1\}$, along the line passing through a particular point $(x^*, y^*) \in [-1,1]^2$ on the xy-plane can be written as $f(z) = F(x^*, y^*, z)$. We explore Fourier representation for such a family of 1D occupancy signals and propose a more compact 2D vector field orthogonal to the view direction, namely *Fourier Occupancy Field (FOF)*, for more efficient representation of the 3D occupancy field, which is detailed in the following subsections.

Fourier Series on a Single Occupancy Line: 1D occupancy signals $\{f(z)\}$ are essentially on-off signals switching at the boundaries of human bodies, which lie in a low-dimensional manifold in the ambient signal space. For accurate

and fast 3D reconstruction, the compact representation should have the following merits:

- *Sampling Scalability*: The representation should be able to adapt to different sampling rates in the inference stage without introducing systematic mismatches except for the inherent approximation error due to sampling. Without retraining, systems with such a representation would support diverse applications with different resolution, quality, and speed requirements.
- *Stable Reconstruction*: The representation should provide stable reconstruction quality for diverse humans with various parameters such as garments, heights, weight, and poses.
- *Low-complexity Reconstruction*: The representation should allow low-complexity reconstruction to support emerging real-time applications such as Holographic Telepresence.

However, most popular representations do not enjoy all the above merits. For instance, discrete Fourier transform and discrete wavelet transform have stable and low-complexity reconstruction, but only support the sampling rate used in preparation of the training set due to their discrete nature. Learned dictionaries are able to provide sparse representations, but require computation-demanding iterative reconstruction, leaving alone the sampling scalability. Without a doubt, sampling scalability is the most difficult property to meet. Pushing the concept of sampling scalability to the limit, one representation would be scalable to arbitrary sampling rates in theory if it allows continuous reconstruction along the z-axis. Note that in practice the representation should stay discrete and compact for efficient inference. *Such a bridge of discrete-representation and continuous-reconstruction motivates us to use Fourier series for the representation of 1D occupancy signals $\{f(z)\}$.* By nature, the Fourier series would have an infinite number of discrete coefficients, which is impractical. Note that most energy of such a 1D occupancy signal concentrates on only a few low-frequency terms, which yields a compact representation by subspace approximation.

Formally, let $f_p(z)$ be periodic extension of $f(z)$. Note that $f_p(z)$ satisfies the Dirichlet conditions:

- $f_p(z)$ is absolutely integrable over one period;
- $f_p(z)$ has a finite number of discontinuities within one period;
- $f_p(z)$ has a finite number of extreme points within one period.

Thus, $f_p(z)$ can be expanded as a convergent Fourier series:

$$f_p(z) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos(n\pi z) + b_n \sin(n\pi z) \right), \quad (2)$$

where $\{a_n\}, \{b_n\} \in \mathbb{R}$ are coefficients of basis functions $\{\cos(nx)\}$ and $\{\sin(nx)\}$, respectively. Note that by defining the occupancy function as Eq. (1), Eq. (2) holds for discontinuity at z^* on the surface S because $f(z^*) = (f(z^*-) + f(z^*+))/2 = 0.5$, where $f(z^*-)$ and $f(z^*+)$ are the left hand limit and right hand limit of $f(z)$ at z^*. For compact

representation, we approximate 1D occupancy signals by a subspace spanned by the first $2N + 1$ basis functions:

$$\hat{f}(z) = \boldsymbol{b}^\top(z)\boldsymbol{c}, \quad (3)$$

where $\boldsymbol{b}(z) = [1/2, \cos(z), \sin(z), \ldots, \cos(Nz), \sin(Nz)]^\top$ is the vector of the first $2N + 1$ basis functions spanning the approximation subspace, and $\boldsymbol{c} = [a_0, a_1, b_1, \ldots, a_N, b_N]^\top$ is the coefficient vector which provides a more compact representation of the 1D occupancy function $f(z)$.

Fourier Occupancy Field: Such a Fourier subspace approximation is applied to all 1D occupancy signals over the xy-plane:

$$\hat{F}(x, y, z) = \boldsymbol{b}^\top(z)\boldsymbol{C}(x, y), \quad (4)$$

where $\boldsymbol{C}(x, y)$ is the $(2N+1)$-dimensional Fourier coefficient vector for the 1D occupancy signal at (x, y). In this way, we obtain the *Fourier Occupancy Field* $\boldsymbol{C} : [-1, 1]^2 \mapsto \mathbb{R}^{2N+1}$ for 3D occupancy field F.

Note that 3D coordinates (x, y, z) in Eq. (4) are continuous and can be sampled at arbitrary rates as needed. Particularly, unlike the discrete Fourier transform or discrete wavelet transform, the subspace dimension for 1D occupancy signals $\{f(z)\}$ is not coupled with the sampling grid of z-axis. This avoids potential sampling mismatches between the sampling grid used in the training data and that in the testing stage. The reconstruction of $\hat{F}$ from the FOF $\boldsymbol{C}$ is sampling scalable in the sense that the sampling rate along the z-axis can be chosen as necessary, *e.g.*, according to the requirement of reconstruction quality and speed. 3D models at any resolution can be readily obtained from the FOF representation by simply altering the sampling rate along the z-axis.

2) FOF with Cosine Series: In FOF-X, we use the cosine series instead of the Fourier series as the basis for the subspace approximation. The cosine series shares similar properties with the Fourier series but works with even periodic extension. Theoretically, even periodic extension can avoid discontinuities, preventing the overshoot known as the Gibbs phenomenon. *Therefore, the cosine series version of FOF representation is used in FOF-X and throughout the following sections of this paper.* The only change is to modify Eq. (2) to:

$$f_{ep}(z) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(n\pi \frac{(z+1)}{2}). \quad (5)$$

Because $f(z)$ is defined on $[-1, 1]$ but not $[0, 1]$, a shift on z is needed, which contributes to the term $\frac{(z+1)}{2}$ in Eq. (5).

B. Image-to-Image Reconstruction Network

In FOF-X, we introduce a novel reconstruction network with dual-sided normal map to mitigate lighting interference in real-world capture scenarios, coupled with integrated SMPL priors that enhance pose robustness under challenging articulations.

1) Reconstruction with Dual-sided Normal Maps: One important change in FOF-X is that we rely on dual-sided normal maps estimated from input RGB images rather than directly using the RGB images themselves to reconstruct human geometry. When testing with RGB input, we found a domain gap between the training images and real captured
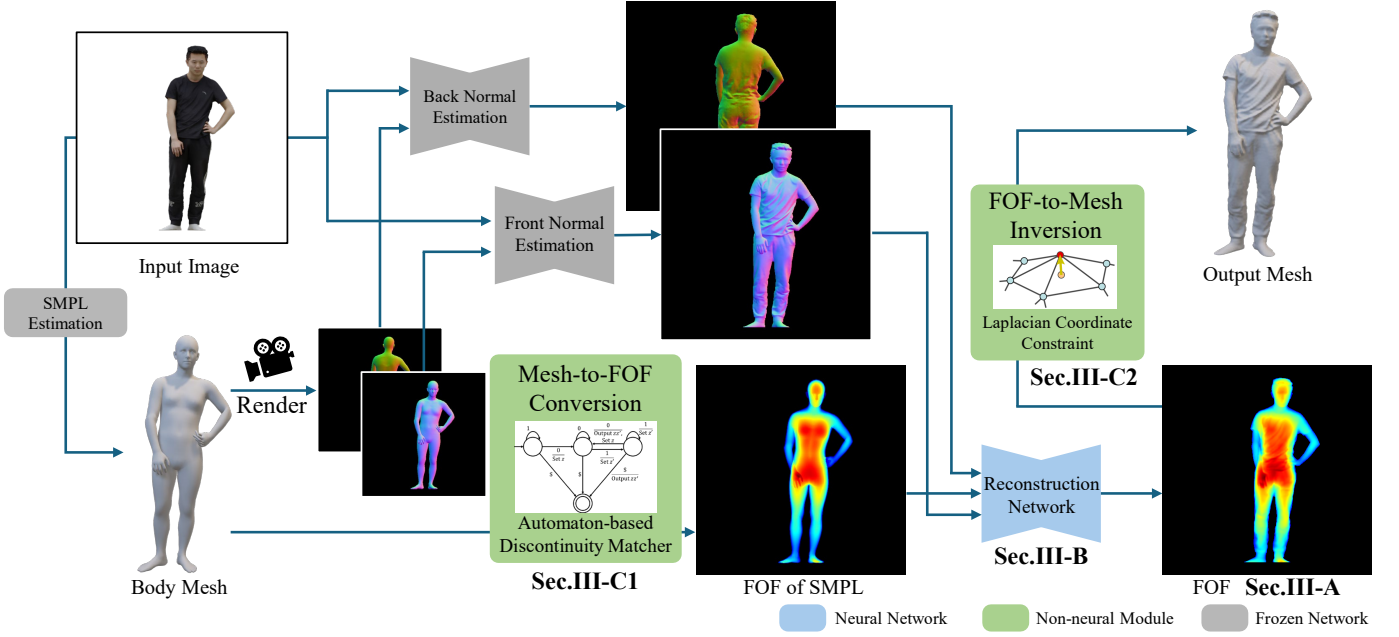
Fig. 3. The overall pipeline of FOF-X for monocular real-time human reconstruction. FOF-X takes an RGB image as input and exploits a SMPL body mesh as a prior with the proposed mesh-to-FOF conversion algorithm (Sec. III-C1), which includes an automaton-based discontinuity matcher to ensure robustness. Based on the rendered SMPL normal maps and input RGB image, the dual-sided normal maps are predicted as the internal representation and decoded to FOF with the SMPL prior through an image-to-image network (Sec. III-B). The FOF representation (Sec. III-A) is then converted to a mesh with the FOF-to-mesh inversion module (Sec. III-C2), incorporating a Laplacian coordinate constraint to enhance the quality of the output mesh.
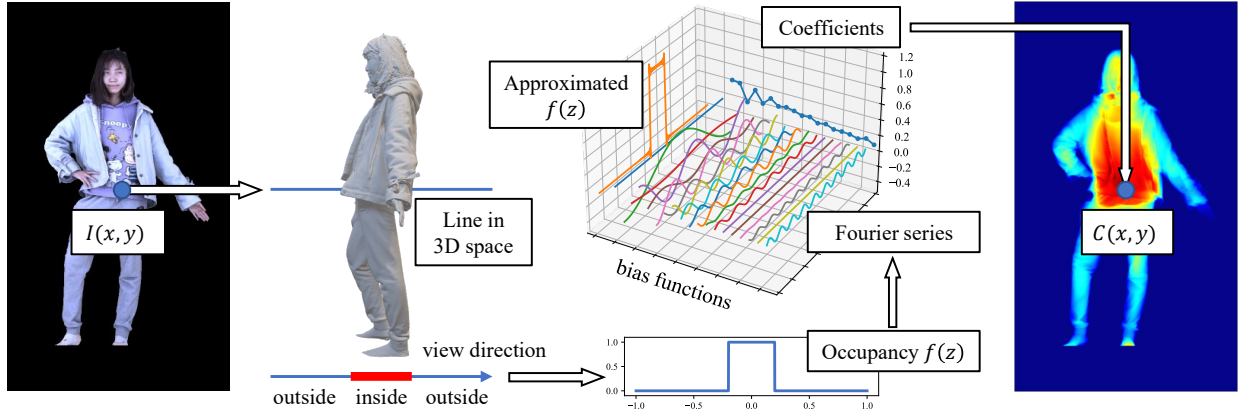

Fig. 4. The proposed Fourier Occupancy Field (FOF). Each pixel on the image I corresponds to a line in 3D space and can be described by an occupancy function $f(z)$, which can be expanded as Fourier series and represented with a vector $C(x, y)$. Therefore, the 3D human model can be encoded as a 2D vector field, forming our FOF. Note that the Fourier series here includes special forms such as the cosine series.

images, which significantly degraded the performance of our method. Although we try to mitigate this gap with a physics-based rendering engine (see Sec. IV-A), the reconstruction network is still sensitive to different textures and lighting conditions. To solve this, we design a new reconstruction framework with the dual-sided normal maps as the internal representation. Together with the FOF of the corresponding SMPL model, it is then fed to the network and used to reconstruct the human geometry. Since normal maps only keep the geometric information and are not affected by texture and lighting, the reconstruction network can focus on geometric information with better robustness and performance.

Thanks to the efficient dimensionality reduction with sub-space approximation, the task of learning FOF from dual-sided normal maps can be essentially regarded as an image-to-image process. To this end, we exploit the image encoder built on the HR-Net [44] framework for its outstanding fitting capability in various vision tasks. We use the weak perspective camera model in our implementation, in which FOF (and thus the reconstructed geometry) is naturally aligned with the input RGB image. Note that FOF can also be used with the normalized device coordinate space (NDC space) so that various camera models, including the perspective camera, are compatible with the framework.

2) SMPL Prior: Reconstructing a 3D object from a single image is a highly ill-posed problem. To address this, 3D priors, such as parametric body models, can be used to enhance the robustness of the model. Thanks to the 2D field nature of

FOF, we propose a novel approach that leverages 3D geometric prior in 2D neural networks. We adopt the SMPL model as a prior and convert it to FOF representation with the proposed algorithm described in Sec. III-C1. Then, we concatenate it with the RGB image as the input together. This can be written as:

$$\boldsymbol{F_H} = G(\boldsymbol{N}_{dual}, \boldsymbol{P}), \tag{6}$$

where $G(\cdot)$ represents an image-to-image translation neural network, which is HR-Net-48 in our implementation, $\boldsymbol{N}_{dual}$ denotes the dual-sided normal maps, and $\boldsymbol{P}$ is the FOF with 16 terms of the corresponding SMPL model estimated from the input image. The use of the prior makes the network more robust to different poses and significantly improves reconstruction quality for cases with high ambiguity.

3) Supervision: To train our network, we sample points on the xy-plane and supervise the coefficients of the lines corresponding to these points with mean squared error (MSE) loss. To make the network more focused on human geometry, we only supervise the human foreground region of the image. The loss function $\mathcal{L}$ is formulated as:

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^{n} \left\| \hat{\boldsymbol{C}}(x_i, y_i) - \boldsymbol{C}(x_i, y_i) \right\|_2^2, \tag{7}$$

where $(x_i, y_i) \in \mathcal{M}$ are sampled foreground pixels, $\mathcal{M}$ is the foreground region, n is the number of foreground pixels, $\boldsymbol{C}(x_i, y_i)$ is the ground-truth FOF coefficient vector, and $\hat{\boldsymbol{C}}(x_i, y_i)$ is the FOF coefficient vector predicted by the network.

C. Inter-Conversion between FOF and Meshes

In FOF-X, the inter-conversion algorithms between FOF and mesh representations are greatly improved.

1) Mesh to FOF with Automaton-based Discontinuity Matcher: To prepare training data, 3D meshes should be converted into FOF as training labels and input priors. By definition Eq. (5), we are to calculate the first few terms of the cosine series for each 1D occupancy signal $f(z)$. For generic signals, this would require numerical integration over a sampled version of $f(z)$, using e.g. the Adaptive Simpson Method, which however may introduce numerical errors and demand considerable computation.

Fortunately, thanks to the particular form of occupancy signals, the cosine coefficients can be derived analytically with exact solutions. Note that the 1D occupancy signal $f(z)$ is actually defined by discontinuities, where the line goes inside or outside the mesh. We extract the discontinuous points of $f(z)$ by designing a rasterizer-like procedure shown in Alg. 1. Formally, suppose that the line associated with $f(z)$ passes though the mesh for k times. For the ith traversing, let z_i and z_i' be the locations going inside and outside the mesh, respectively. Then, the set $\mathcal{Z} \triangleq \{(z_i, z_i')\}_{i=1}^k$ collects those intervals of the line that are inside the mesh. According to the definition of the occupancy signal, we have: $f(z) = 1, z \in \mathcal{Z}$, and $f(z) = 0$ otherwise, except for a finite number of

---

Algorithm 1 Mesh-to-FOF conversion

Input: A triangle mesh $M = \{t_i\}_{i=1}^n$ ▷ A set of triangles
Output: FOF $C \in \mathbb{R}^{H \times W \times N}$
1: $B[x, y] \leftarrow \varnothing, \forall (x, y) \in \{1, ..., H\} \times \{1, ..., W\}$
2:   ▷ Initialize buffers
3: $C \leftarrow 0$ ▷ Initialize FOF
4: **for** $i = 1$ to n **do**   ▷ parallelized
5:      Compute the bounding box b_i of t_i
6:      Compute the normal direction n_i of t_i
7:      **for** each pixel (x, y) lying inside b_i **do**
8:          **if** the pixel (x, y) is inside t_i **then**
9:             Compute the depth z of (x, y)
10:            **if** n_i faces front **then**
11:                $B[x, y] \leftarrow B[x, y] \cup (z, 0)$ ▷ Add to buffer
12:            **else**
13:                $B[x, y] \leftarrow B[x, y] \cup (z, 1)$ ▷ Add to buffer
14:            **end if**
15:          **end if**
16:      **end for**
17: **end for**
18: **for** $(x, y) \in \{1, ..., H\} \times \{1, ..., W\}$ **do**   ▷ parallelized
19:      Sort the discontinuities in the buffer $B[x, y]$
20:      Apply the automaton to the buffer $B[x, y]$
21:      Compute integral results $\mathcal{I}$ with the buffer $B[x, y]$
22:      $C[x, y] \leftarrow \mathcal{I}$
23: **end for**

---

discontinuities with measure zero. Therefore, we have the following analytical solution for the cosine coefficients:

$$\begin{aligned}
a_n &= \int_{-1}^{1} f(z) \cos(t_n(z + 1)) dz \\
&= \frac{1}{t_n} \sum_{i=1}^{k} \sin(t_n(z_i' + 1)) - \sin(t_n(z_i + 1)),
\end{aligned} \tag{8}$$

where $t_n = \frac{n\pi}{2}$. The FOF representation is obtained by calculating the cosine coefficients for all the 1D occupancy signals. Since the number of intervals k is generally small, *e.g.*, 1~3 for human meshes, the calculation of a_n is of $O(1)$ complexity. Thus, the transformation from discontinuities to FOF, shown in Alg. 1, has a computational complexity of $O(HWN)$, and can be parallelized on GPUs for real-time running, where H and W are the height and width of the FOF, respectively.

However, due to numerical errors and some non-watertight meshes, rasterization may produce unpaired $\mathcal{Z}$, i.e., a line may encounter successive discontinuities both going out of or into the mesh. And it can come to a going outside point first instead of a going inside point. To solve this, we propose an automaton-based discontinuity matcher to enhance the robustness. Unlike what we do in the conference version [19], we collect those discontinuities and normal vectors for each line first and process them with an automaton before working out the integral. Let's consider the discontinuities corresponding to one pixel. We sort these points by z first and then represent them with a sequence of 0s and 1s, where 0 stands for the normal vector facing front and 1 stands for the normal vector
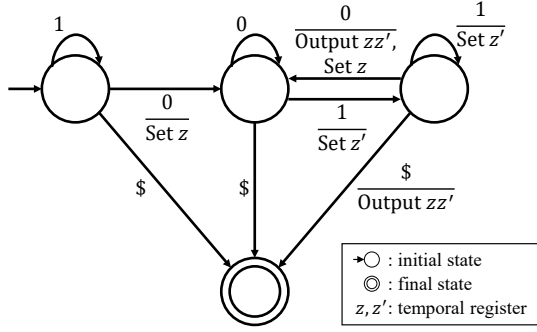
Fig. 5. Automaton-based discontinuity matcher used to process discontinuities on each pixel. The results of with and without Automaton-based discontinuity matcher are shown in Fig. 2.
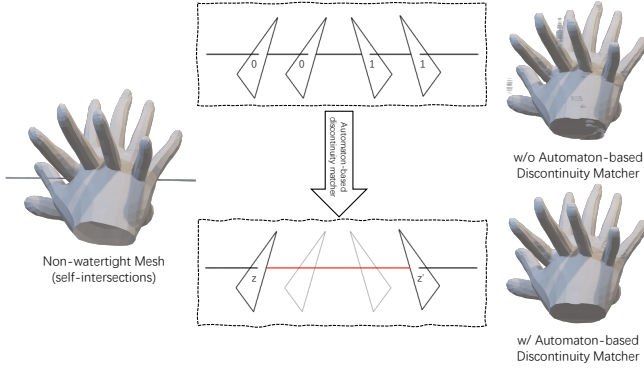

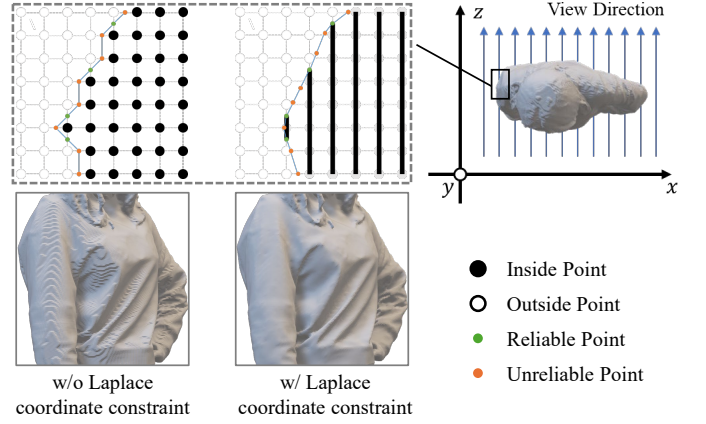Fig. 6. Example of Discontinuity Point Matching via Automaton.

Fig. 7. Our FOF-to-mesh algorithm with Laplacian coordinate constraint. Compared with the original Marching Cubes, our conversion algorithm avoids the stair-step artifacts.

facing backwards. We add a stop letter $ to the sequence and then process it with the automaton shown in Fig. 5, which effectively handles irregular patterns (initial 1s and repeating 0s or 1s). After being filtered by the automaton, the produced sequence can then be integrated with Eq. (8). For watertight meshes, both our automaton and the conference version produce identical results, correctly converting meshes to FOF. For non-watertight meshes (e.g., surfaces with open boundaries or self-intersections, partly missing, or with non-manifold connectivity), the conference version method fails to compute FOF accurately, leading to floating artifacts (Fig. 2). Our automaton-based discontinuity matcher robustly handles irregular patterns arising from non-watertight meshes (including holes, self-intersections, and duplicate faces), effectively preventing reconstruction artifacts caused by such topological imperfections. Fig. 6 shows an example of our automaton processing a non-watertight mesh.

2) FOF to Mesh with Laplacian Coordinate Constraint:
Extracting meshes is necessary for applications such as mesh-based rendering and animation. Recent methods based on implicit neural representations need to apply an MLP network on a 3D sampling grid. Instead, as described in Eq. (4), the reconstruction of 3D occupancy field from FOF is simply a multiplication of two tensors. Then, the 3D mesh is extracted from the iso-surface of the 3D occupancy at the threshold of 0.5 with the Marching Cubes algorithm [45]. Both steps can be parallelized on GPUs for fast mesh generation.

However, it is crucial to note that the occupancy field generated by FOF exhibits a view-dependent bias, resulting in observable stair-step artifacts on the reconstructed mesh. This bias marginally impacts the geometric accuracy and only detracts from the visual quality. As illustrated in Fig. 7, when applying Marching Cubes on FOF, only the points on the edges parallel to the z-axis remain precise. To address this, we propose a Laplacian coordinate constraint to eliminate the view bias. We adapt the Marching Cubes algorithm to incorporate an additional attribute for each point, which indicates the reliability of the point's coordinates. The points produced on the edges parallel to the z-axis are reliable, and the points produced on the edges orthogonal to the z-axis are not. For those unreliable points, their coordinates are adjusted to confer greater mesh smoothness. This refinement is accomplished by minimizing the mean square of the Laplacian coordinates, expressed as:

$$\underset{X_i, i \in U}{\arg\min} \|(D - A)X\|_2^2, \qquad (9)$$

where D is a diagonal matrix with $D_{ii} = d_i$ on the diagonal, d_i is the number of neighbors of the ith point, $X \in \mathbb{R}^{n \times 3}$ is the coordinate matrix to be optimized, and A is the adjacency matrix. $A_{ij} = 1$ if points v_i and v_j are adjacent, otherwise $A_{ij} = 0$. U is the index set of unreliable points, which are to be optimized. Unlike excessive Laplacian smoothing iterations that degrade geometric features into spherical approximations, our Laplacian coordinate constraint preserves topological fidelity and guarantees convergence to ground-truth mesh shapes.

IV. EXPERIMENTS

A. Datasets

We collect THuman2.1 [47], CustomHumans [48], and CAPE [46] for our experiments. To maintain reproducibility, all the datasets we use are publicly accessible. For fair comparisons, we follow ICON [3] by training our method on THuman2.0, which is a subset of the first 526 items in THuman2.1, and using the CAPE dataset for evaluation. Additionally, we create two additional benchmarks on THuman2.1 and CustomHumans to evaluate our method and all the baselines.

TABLE II
QUANTITATIVE COMPARISON WITH EXISTING METHODS.

	Year	Time↓	CAPE [46]			THuman2.1 [47]			CustomHumans [48]		
			Chamfer↓	P2S↓	Normal↓	Chamfer↓	P2S↓	Normal↓	Chamfer↓	P2S↓	Normal↓
PIFu* [1]	2019	9.98	1.911	2.265	1.461	3.589	3.792	2.504	3.635	3.932	2.365
PIFuHD [11]	2020	15.90	3.279	3.393	1.903	3.963	3.986	2.620	4.184	4.115	2.432
PaMIR* [12]	2021	36.48	1.426	1.629	1.377	1.549	1.771	1.720	1.211	1.404	1.350
ICON [3]	2022	35.00	0.967	1.001	0.909	0.968	0.881	1.338	0.798	0.755	1.039
ICON$_{fl}$ [3]	2022	35.56	0.817	0.814	0.755	0.948	0.860	1.254	0.811	0.761	0.984
D-IF [30]	2023	61.15	0.758	0.714	0.752	1.210	1.082	1.332	1.043	0.955	1.098
ECON$_{if}$ [4]	2023	124.55	0.947	0.956	0.928	1.186	1.224	1.442	1.015	1.027	1.106
ECON$_{ex}$ [4]	2023	108.99	0.887	0.864	0.844	1.094	1.049	1.301	1.055	1.040	1.091
GTA [29]	2023	50.29	0.738	0.761	0.808	1.514	1.524	1.555	1.382	1.435	1.332
SiFU [2]	2024	62.38	**0.676**	0.677	0.733	1.015	1.004	1.271	0.880	0.900	0.994
SiTH [18]	2024	83.18	1.016	1.030	1.051	1.218	1.160	1.469	1.111	1.061	1.216
PSHuman [43]	2025	76.69	2.208	1.922	0.718	1.464	1.276	**0.675**	2.066	1.815	<u>0.772</u>
FOF-Base† [19]	2022	<u>0.09</u>	4.583	3.825	2.345	3.597	3.408	2.489	4.677	4.084	2.671
FOF-SMPL† [19]	2022	-	0.765	0.709	0.649	0.959	0.849	1.120	0.850	0.759	0.882
FOF-X(HR32) (Ours)	-	**0.02**	<u>0.708</u>	**0.654**	**0.634**	<u>0.858</u>	<u>0.768</u>	1.049	<u>0.731</u>	<u>0.663</u>	0.783
FOF-X(HR48) (Ours)	-	**0.02**	0.716	<u>0.672</u>	<u>0.639</u>	**0.835**	**0.754**	<u>1.033</u>	**0.718**	**0.658**	**0.771**

To obtain RGB images, we render the meshes in Blender using 630 HDRi environment texture maps from Poly Haven[1]. The HDRi maps serve both as panoramic background images and sources of ambient lighting. Each subject in the benchmarks is rendered to four images, uniformly distributed along the yaw axis with a fixed elevation of 0 degrees. For consistency with the CAPE dataset, all the meshes are normalized to a height of 1.8 meters. In practice, we find that the human body occupying a large percentage of the image and a clearer image typically lead to better reconstruction results.

B. Implementation Details

The reconstruction network of FOF-X takes dual-sided normal maps and the FOF of SMPL as input. The dimensions of the input are $512 \times 512 \times (3+3+16)$, and the dimensions of the produced feature map are $128 \times 128 \times 256$. We adopt HRNet-W48 [49] as the image encoder of our system due to its multi-scale perception ability. Since the SMPL mesh only provides a rough shape prior, the high-frequency components corresponding to geometric details are redundant for the input. Empirically, we find that the first 16 terms of the SMPL mesh is sufficient for the shape prior. The decoder progressively upsamples the feature map to 256×256 and 512×512 using bilinear interpolations and residual convolutional blocks, while retaining a channel number of 256 throughout the process. Finally, a convolutional layer projects the features to a FOF with a resolution of $512 \times 512 \times 128$. To cope with single view image as input, we use the off-the-shelf image-to-image translation network used in ICON [3] for generating dual-sided normal maps from the single image.

During training, generating multi-view normal maps and FOFs requires a considerable amount of storage and forwarding these data to GPUs can significantly slow down the training process. However, thanks to our fast FOF conversion algorithm implemented as a PyTorch CUDA extension, we can efficiently render meshes to FOF and normal maps. This allows us to keep

all meshes in memory and generate training data on the GPU in real time. Additionally, this approach enables random view sampling during training, removing the constraints of training on fixed views.

We utilize the Adam optimizer [50] and set the learning rate to 2×10^{-5} without decay. Our methods and all the variants are trained on a single RTX-4090 GPU using automatic mixed-precision training with a batch size of 8. Convergence takes approximately 10 hours on the THuman2.0 dataset and about 2 days on the THuman2.1 training set.

For consistency with previous works [1], [3], [4], we employ a weak perspective camera. The scale of the weak perspective camera is fixed at 1 for training and evaluation, resulting in an orthogonal projection. But note that we can transform world space to NDC (Normalized Device Coordinates) space to support perspective cameras straightforwardly. We follow the post-processing for hands described in ECON [4], fusing the hands of the parametric model with the reconstructed body. This strategy is not used in our real-time systems.

C. Metrics

We use Chamfer distance, P2S (point-to-surface) distance, 3D IoU (Intersection over Union), and normal map error for quantitative evaluation. In the following, we provide a summary of metrics used in our experiments:

a) Chamfer and P2S Distances: We use Chamfer and P2S distances to evaluate the surface geometric error between the predicted and ground-truth meshes. Here, the Chamfer distance d_{cd} is defined as:

$$d_{cd}(M_{pr}, S_{pr}, M_{gt}, S_{gt}) = \frac{1}{2|S_{pr}|} \sum_{x \in P_{pr}} d(x, M_{gt}) + \frac{1}{2|S_{gt}|} \sum_{x \in S_{gt}} d(x, M_{pr}), \quad (10)$$

where S_{pr}, S_{gt} are the sets of points uniformly sampled on the predicted and ground-truth mesh surfaces M_{pr} and M_{gt} separately, and $d(x, M)$ is the minimal distance from a point
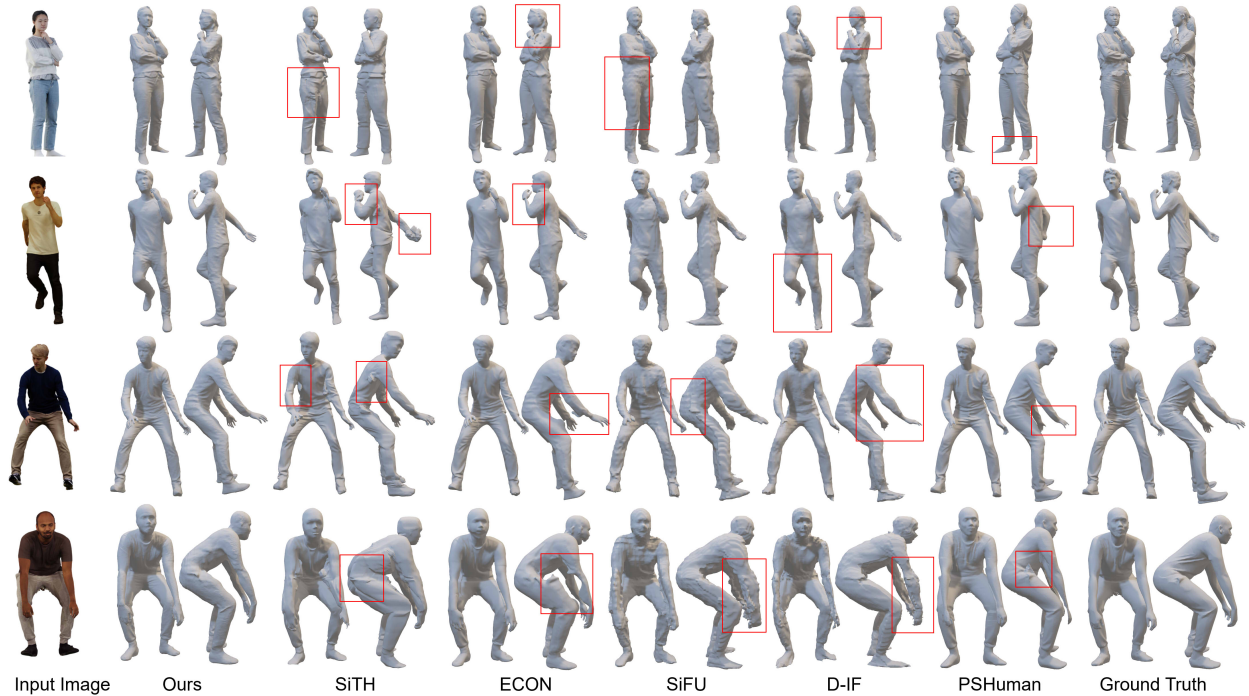
Fig. 8. Qualitative comparison against the state-of-the-art methods.

x to the mesh surface M. We report the averaged Chamfer distances of all test samples. We also measure the average point-to-surface distance from the points on the reconstructed surface to the ground-truth mesh. The P2S distance d_{p2s} is defined as:

$$d_{p2s}(S_{pr}, M_{gt}) = \frac{1}{|S_{pr}|} \sum_{x \in P_{pr}} d(x, M_{gt}). \quad (11)$$

Note that this metric alone does not fully measure geometric accuracy. For instance, the P2S distance between a ground-truth mesh and a partial ground-truth mesh would be zero. However, we also report this metric to align with the existing published works.

b) Normal Difference: To measure the visual quality of the reconstructed geometries, we render them as normal maps and compute the mean squared error with the ground-truth normal maps. For each reconstructed mesh, we render four normal maps from views at $\{0°, 90°, 180°, 270°\}$ based on vertex normals. We do not apply any anti-aliasing techniques when rendering these normal maps. Such a metric provides an effective assessment of the visual fidelity of the reconstructions.

D. Comparison with the State of the Arts

We compare our method with the state-of-the-art approaches, including PIFu [1], PIFuHD [11], PaMIR [12], ICON [3], ECON [4], D-IF [30], GTA [29], SiFU [2], SiTH [18], and PSHuman [43] on CAPE [46], THuman2.1 [47], and CustomHumans [48].

a) Quantitative Results: The quantitative results are summarized in Table II. * denotes the re-implemented version in

ICON [3]. † denotes the re-implemented version with the cosine series and the improved inter-conversion algorithms. FOF-X(HR32) and FOF-X(HR48) are FOF-X model with HRNet-W32 and HRNet-W48 respectively. All methods listed, except PIFuHD [11] and PSHuman [43], are trained on THuman2.0 [47]. It is worth noting that PSHuman's [43] dataset includes THuman 2.1. For PIFu [1] and PaMIR [12], similar to ICON [3], we use the re-implemented versions, as the official releases are trained with additional commercial datasets. Since some methods do not provide the training sample lists, we use THuman2.1 instead of THuman2.0 to evaluate all methods. We use $512 \times 512 \times 512$ resolution for the input of all these methods. We do not compare our method with [13] and [51] because their codes are not released. For simplicity, we also ignore some works that have already been compared with the state-of-the-art methods listed above. We found that some methods are sensitive to the ratio of the human subject's size in the image. To ensure consistency, we use a ratio of 0.9 for all methods, corresponding to the 1.8m height normalization mentioned in the datasets section. Ground-truth SMPL meshes are used to better demonstrate the capability of the evaluated methods.

As can be seen, our methods show the lowest reconstruction errors compared to other methods on CustomHumans datasets. As mentioned in [18], CAPE provides the fitted SMPL+D meshes instead of the original scanned meshes as the ground-truth, leading to mismatches between the image details and the ground-truth meshes. This may cause a potential bias in evaluation. Despite this, our FOF-X can still achieve a comparable metric on CAPE. In addition, on the THuman2.1 dataset, we are only slightly inferior to PSHuman [43] in terms of normals. However, we note that PSHuman's [43]
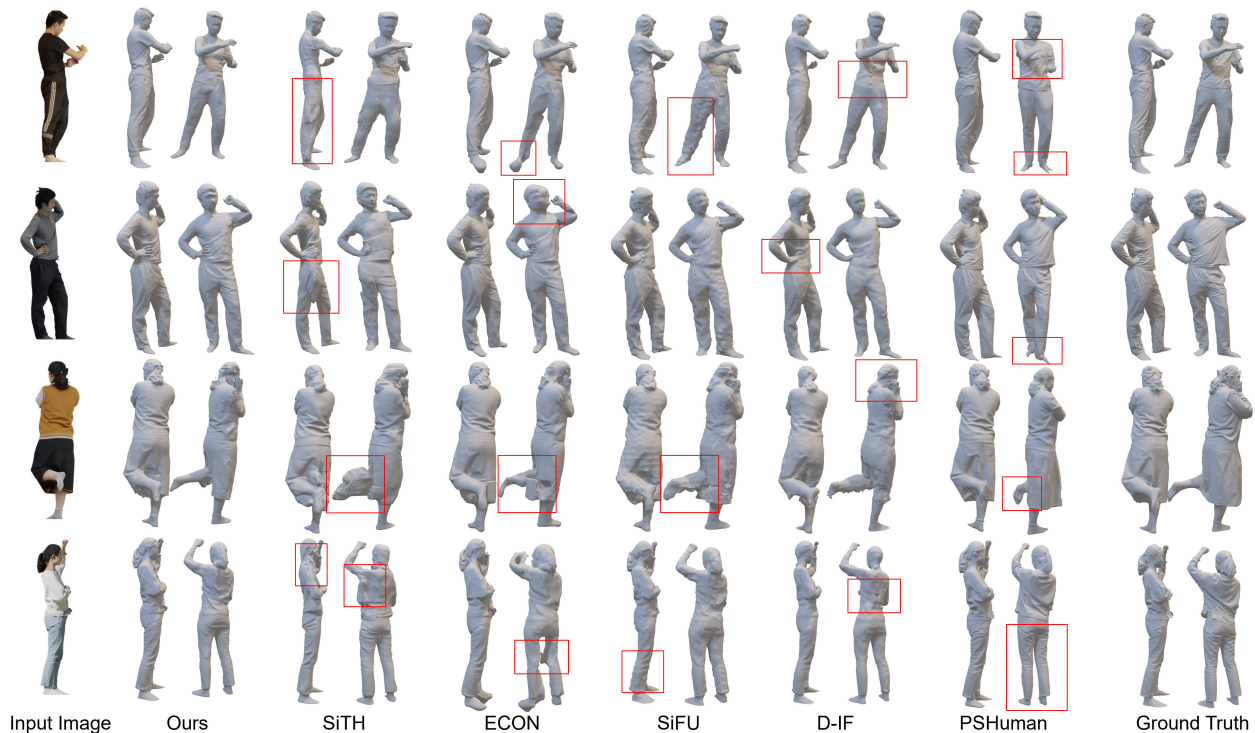
Fig. 9. Qualitative comparison against the state-of-the-art methods on non-frontal view inputs.

training dataset incorporates THuman2.1 without providing clear train/test splits, potentially allowing some test data to be included in training. Notably, our method demonstrates robustness across all benchmarks, achieving better or comparable performance to other computationally heavy methods.

b) Qualitative Results: We conduct a qualitative comparison in Fig. 8. As can be seen, SiTH [18] and PSHuman [43] generally produces good results but sometimes introduces artifacts in the limbs and side shapes. ECON [4] suffers from self-occlusions which is the essential limitation of its optimization-based solution; SiFU [2] cannot recover detailed surfaces, although it can produce accurate rough shapes; D-IF [30] has difficulties dealing with limb shapes due to the way it uses the SMPL prior. In contrast, our method is able to reconstruct plausible 3D human models under different body poses and shapes. In terms of surface quality and geometric accuracy, our method outperforms other state-of-the-art approaches. Notably, our method is much more lightweight than others, allowing for real-time reconstruction.

Fig. 9 shows additional comparison on non-frontal view inputs. Generally, it is more challenging to obtain information about the body and limb shape from non-frontal view inputs. In this case, the results of SiTH [18], D-IF [30], and ECON [4] all experience varying degrees of performance degradation. SiFU [2] is still constrained by the resolution issue in producing detailed surfaces. PSHuman [43] has certain difficulties in processing limb shapes. In contrast, our method is more robust and still achieves good results.

c) Comparison of Running Times: Table II shows the comparison results in terms of running time. All methods except PSHuman [43] are tested on a single RTX-3090 GPU, where FOF-X represents the acceleration of FOF-X using TensorRT. PSHuman [43] is tested on the A100 GPU due to memory limitations. Our FOF-X takes 0.02 seconds per frame, demonstrating 500× acceleration over the fastest baseline (PIFu [1]: 9.98s). The original system presented in our conference publication, FOF [19], is implemented in Python. Due to limitations in the original mesh-to-FOF algorithm, the SMPL prior is not supported in the FOF real-time system. Since FOF-SMPL requires the execution of three separate scripts, its running time is not reported. Our enhanced pipeline FOF-X adds additional components such as SMPL estimation, image segmentation, and rendering while maintaining real-time performance. Notably, all variants of our method (Python/TensorRT) consistently outperform existing works in terms of efficiency by two orders of magnitude.

E. Ablation Study

In this section, we conduct ablation experiments to validate the effectiveness of the settings and the proposed components in our FOF representation and FOF-X framework. Note that the ablation studies on the number of Fourier terms to use and the noise robustness of FOF are not using neural networks. These experiments are designed to explore the intrinsic properties of the FOF representation itself rather than the performance of the reconstruction algorithm.

1) Number of Terms to Use: In the original FOF, we use $2N + 1 = 31$ terms of the Fourier series to represent the occupancy field, which has been shown to be sufficient for maintaining geometric accuracy. However, we observe that

TABLE III
GEOMETRIC METRICS FOR DIFFERENT NUMBERS OF FOF COMPONENTS.

N	8	16	32	64	128	256
P2S	1.342	0.466	0.148	0.054	0.027	0.024
Chamfer	2.544	0.529	0.168	0.062	0.030	0.025

$N = 8$ $\quad$ $N = 16$ $\quad$ $N = 32$ $\quad$ $N = 64$ $\quad$ $N = 128$ $\quad$ $N = 256$ $\quad$ Ground Truth
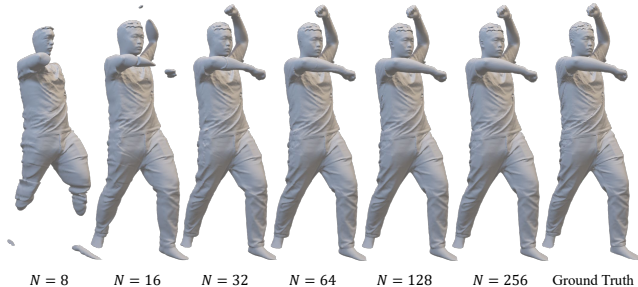
Fig. 10. Results of different number N of coefficients used in FOF-X.

TABLE IV
GEOMETRIC METRICS FOR DIFFERENT RESOLUTION OF FOF.

Res	16	32	64	128	256	512
P2S	3.580	1.764	0.885	0.396	0.139	0.028
Chamfer	3.655	1.746	0.889	0.403	0.146	0.031

Res=16 $\quad$ Res=32 $\quad$ Res=64 $\quad$ Res=128 $\quad$ Res=256 $\quad$ Res=512 $\quad$ Ground Truth

Fig. 11. Results of different resolution Res used in FOF-X. We use $Res = 256$ in our implementation.

the absence of high-frequency components can lead to visual artifacts. Therefore, we use $N = 128$ terms of the cosine series in FOF-X. We illustrate the impact of varying N on the accuracy of the approximation in Fig. 10. The metrics for different values of N are also presented in Table III. When $N \leq 16$, noticeable artifacts appear on the meshes. Noting the geometric details in the ear region, although using $N = 32$ preserves most geometric information, we prefer $N = 128$ in FOF-X to ensure higher visual quality and eliminate potential artifacts.

2) Sampling Scalability: The frequency-domain resolution of FOF can be adaptively inferred according to the time needs of the system, without the need for retraining. We explored the effect of resolution (number of components) on the results taking $N = 128$. Fig. 11 illustrates the effect of changing resolution on the accuracy of the results. Metrics of different N values are also given in Table IV.

3) Noise Robustness: Neural networks often struggle to produce precise results, and their outputs can be considered as samples from a certain distribution, leading to inherent noise in the results. The geometric representation used must therefore be robust to noise and small perturbations. Therefore, representations that are sensitive to noise are not suitable for machine learning. To verify the noise robustness of FOF, we manually add relative Gaussian noise with different variances and show the quantitative results in Fig. 12 and Table V. Based on these experiments, we can conclude that the FOF is very robust to noise and can keep a general shape even with large noise. Thus, it is reasonable and feasible to use neural networks to estimate FOFs.

4) Dual-sided Normal Maps: To mitigate the performance degradation caused by the domain gap between the training images and the real images, we take the RGB images as input and first map them to dual-sided normal maps for the subsequent pipeline. To verify the effectiveness of such a strategy, we compare FOF-X with our baseline method, which uses RGB images as input directly. As shown in Table II, with dual-sided normal maps as an internal representation, FOF-X-32 is more robust than the original FOF (FOF-SMPL†) on test
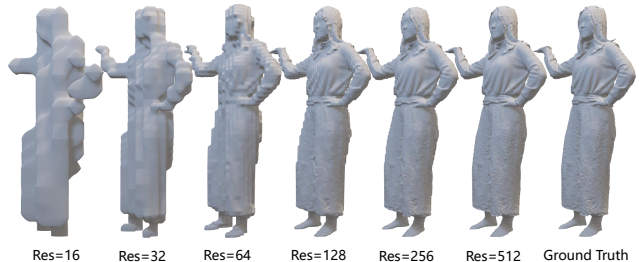
sets. Note that FOF-X is trained on ground-truth normal maps and tested on maps generated by an off-the-shelf model. Fig. 1 shows the visual results.

5) Improved Inter-Conversion Algorithm: Fig. 2 shows the effectiveness and robustness of our improved inter-conversion algorithm. As can be seen, our parallelized mesh-to-FOF algorithm demonstrates strong robustness. The quantitative results for two mesh-to-FOF algorithms are also given in Table VI. With the automaton-based discontinuity matcher, it effectively avoids the floating masses caused by duplicate or missing discontinuities. This algorithm is integrated into the real-time pipeline for its excellent parallelizability. Table VII shows the quantitative results of the FOF-to-mesh methods. To better represent the visual effect of the mesh, we render the mesh as a normal diagram and introduce visual metrics[2] to evaluate results. Our improved FOF-to-mesh algorithm avoids the stair-step artifacts caused by view bias effectively. Although the Laplacian coordinate constraint solution is a little time-consuming, we can use Laplacian smoothing in the real-time pipeline as an alternative and apply our full FOF-to-mesh algorithm during the mesh export stage.

F. The Real-time Pipeline

To perform real-time monocular human reconstruction, we design a four-stage pipeline. In the first stage, we get an image from the video stream with OpenCV [52], use RVM [53] to get a human mask to remove the background, and predict a SMPL model as prior with 4D-humans [54]. RVM is originally a matting method, and we use a threshold of 0.5 to obtain a binary mask. In the second stage, a CUDA-implemented rasterizer is used to render dual-sided normal maps and a FOF representation for the predicted SMPL. The proposed mesh-to-FOF conversion algorithm is integrated into the rasterizer. In the third stage, the pre-trained dual-sided normal estimator and our reconstruction network are encapsulated into a single

---

[2]SSIM (structural similarity), PSNR (peak signal-to-noise ratio) and LPIPS (learned perceptual image patch similarity) measure structural preservation, pixel-level accuracy and human perceptual consistency, respectively.

TABLE V
GEOMETRIC METRICS FOR DIFFERENT NOISE LEVEL

Noise level	0%	5%	10%	15%	20%	25%	30%
P2S	0.027	0.030	0.035	0.041	0.048	0.075	0.224
Chamfer	0.030	0.032	0.037	0.043	0.049	0.065	0.141

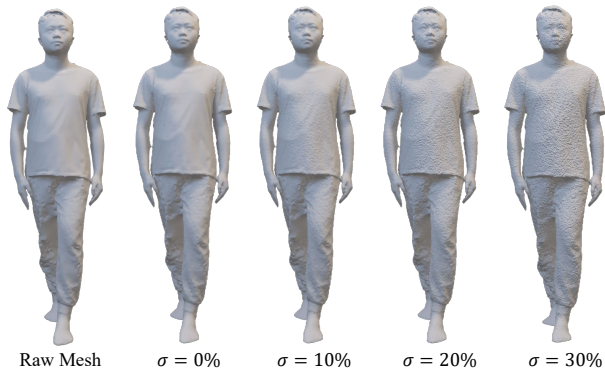Raw Mesh $\sigma = 0\%$ $\sigma = 10\%$ $\sigma = 20\%$ $\sigma = 30\%$

Fig. 12. Results showing the effects of different noise levels applied to FOF-X, demonstrating FOF-X's robustness to noise.

Fig. 13. Our system applied to perform real-time 3D human reconstruction from a single monocular RGB input. Please see the supplement video for the live demonstration.

TABLE VI
GEOMETRIC METRICS FOR DIFFERENT MESH-TO-FOF ALGORITHMS

	Chamfer↓	P2S↓	Normal↓
Algorithms Used in FOF	0.447	0.206	0.264
Algorithms Used in FOF-X	**0.041**	**0.063**	**0.099**

TABLE VII
VISUAL METRICS FOR DIFFERENT FOF-TO-MESH ALGORITHMS

	SSIM↑	PSNR↑	LPIPS↓
w/o Laplacian coordinate constraint	0.959	32.09	0.018
w/ Laplacian coordinate constraint	**0.965**	**32.35**	**0.015**

additional improvements to the conversion algorithm and the reconstruction algorithm. For the conversion algorithm, an automaton-based discontinuity matcher is designed to convert a triangle mesh to a FOF efficiently and robustly. Thanks to this, the SMPL mesh prior can be added into the real-time pipeline, further enhancing the performance of the system. Besides, the FOF to mesh conversion is also improved with the Laplacian coordinate constraint, eliminating the stair-step artifacts. For the reconstruction framework, we predict dual-sided normal maps from input RGB images first and then use dual-sided normal maps instead of RGB images as the input of the main network, which greatly mitigates the performance degradation caused by the domain gap between the training images and the real images. The efficacy of FOF and FOF-X is validated on 3D scanned and in-the-wild datasets, where our approaches effectively recover detailed human geometry with better quality than previous solutions.

Limitations and Future Work. The FOF representation is based on the Fourier series expansion of the square-wave-like function. When applying FOF to thin objects, the spectrum of the function will contain many high-frequency components. In such cases, we need more terms in the Fourier series to approximate the function. The number of terms we need is roughly inversely proportional to the thickness of the object. In our experiments, we found that FOF fails to reconstruct objects that are too thin, like single layer mesh. Although wavelet transformation and short-time Fourier transformation seem to be a promising option, the implementation and computational efficiency still need more exploration. In addition, the FOF representation shows significant potential for broader applications. Its mathematical formulation and efficient parameterization could be extended to general object representation and even scene-level modeling through appropriate architectural modifications. These promising directions will be explored in future work to enhance the framework's capabilities and applicability.

neural network module, which is used to reconstruct the FOF. In the final stage, the FOF is resized to a proper resolution (256×256 in our implementation) and transformed into a mesh. All four stages are implemented in PyTorch and accelerated with TensorRT, running on a single RTX-4090 GPU. Our pipeline works at 30+ FPS, which is limited by the camera (Logic C920 Pro). With a recorded video, it can achieve a frame rate of about 50 FPS. Fig. 13 shows the real-time reconstruction system (all human subjects provided informed consent for data collection). Please check the demo video for more results.

V. CONCLUSION AND DISCUSSION

Conclusion. In this paper, we first present the Fourier Occupancy Field (FOF) representation for monocular real-time 3D human reconstruction and further extend it to the FOF-X framework for more detailed results. FOF is an efficient and flexible 3D representation defined as a 2D vector field orthogonal to the z-axis. When extending the FOF representation to the FOF-X reconstruction framework, we introduce

VI. ACKNOWLEDGEMENTS

REFERENCES

[1] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, "PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization," in *Proc. ICCV*, 2019.

[2] Z. Zhang, Z. Yang, and Y. Yang, "SIFU: Side-view conditioned implicit function for real-world usable clothed human reconstruction," in *Proc. CVPR*, 2024.

[3] Y. Xiu, J. Yang, D. Tzionas, and M. J. Black, "ICON: Implicit Clothed humans Obtained from Normals," in *Proc. CVPR*, 2022.

[4] Y. Xiu, J. Yang, X. Cao, D. Tzionas, and M. J. Black, "ECON: Explicit clothed humans optimized via normal integration," in *Proc. CVPR*, 2023.

[5] A. A. Alatan, Y. Yemez, U. Gudukbay, X. Zabulis, K. Muller, Ç. E. Erdem, C. Weigel, and A. Smolic, "Scene representation technologies for 3DTV—A survey," *IEEE Trans. Circuits and Systems for Video Technology*, 2007.

[6] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu, "DeepHuman: 3D human reconstruction from a single image," in *Proc. ICCV*, 2019.

[7] H. Zhu, X. Zuo, H. Yang, S. Wang, X. Cao, and R. Yang, "Detailed avatar recovery from single image," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2021.

[8] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor, "Tex2shape: Detailed full human body geometry from a single image," in *Proc. ICCV*, 2019.

[9] K. Li, H. Wen, Q. Feng, Y. Zhang, X. Li, J. Huang, C. Yuan, Y.-K. Lai, and Y. Liu, "Image-guided human reconstruction via multi-scale graph transformation networks," *IEEE Trans. Image Processing*, 2021.

[10] Y. Huang, H. Yi, Y. Xiu, T. Liao, J. Tang, D. Cai, and J. Thies, "TeCH: Text-guided Reconstruction of Lifelike Clothed Humans," in *Proc. 3DV*, 2024.

[11] S. Saito, T. Simon, J. Saragih, and H. Joo, "PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization," in *Proc. CVPR*, 2020.

[12] Z. Zheng, T. Yu, Y. Liu, and Q. Dai, "PaMIR: Parametric model-conditioned implicit representation for image-based human reconstruction," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2021.

[13] T. Alldieck, M. Zanfir, and C. Sminchisescu, "Photorealistic monocular 3D reconstruction of humans wearing clothing," in *Proc. CVPR*, 2022.

[14] R. Li, Y. Xiu, S. Saito, Z. Huang, K. Olszewski, and H. Li, "Monocular real-time volumetric performance capture," in *Proc. ECCV*, 2020.

[15] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graphics*, 2015.

[16] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3D hands, face, and body from a single image," in *Proc. CVPR*, 2019.

[17] A. A. A. Osman, T. Bolkart, and M. J. Black, "STAR: A sparse trained articulated human body regressor," in *Proc. ECCV*, 2020.

[18] H.-I. Ho, J. Song, and O. Hilliges, "SiTH: Single-view textured human reconstruction with image-conditioned diffusion," in *Proc. CVPR*, 2024.

[19] Q. Feng, Y. Liu, Y.-K. Lai, J. Yang, and K. Li, "FOF: Learning fourier occupancy field for monocular real-time human reconstruction," in *Proc. NeurIPS*, 2022.

[20] M. Zhang, Q. Feng, Z. Su, C. Wen, Z. Xue, and K. Li, "Joint2Human: High-quality 3D human generation via compact spherical embedding of 3D joints," in *Proc. CVPR*, 2024.

[21] Y. Li, F. Luo, and C. Xiao, "Diffusion-FOF: Single-view clothed human reconstruction via diffusion-based Fourier occupancy field," in *Proc. CVPR*, 2024.

[22] Y. Yang, Q. Feng, Y.-K. Lai, and K. Li, "R²Human: Real-time 3D human appearance rendering from a single image," in *Proc. ISMAR*, 2024.

[23] Y. Yang, Q. Feng, Y.-K. Lai, and K. Li, "Real-time 3D human reconstruction and rendering system from a single rgb camera," in *SIGGRAPH Asia 2024 Technical Communications*, 2024.

[24] H. Zhang, Y. Tian, X. Zhou, W. Ouyang, Y. Liu, L. Wang, and Z. Sun, "PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop," in *Proc. ICCV*, 2021.

[25] M. Kocabas, C.-H. P. Huang, O. Hilliges, and M. J. Black, "PARE: Part attention regressor for 3D human body estimation," in *Proc. ICCV*, 2021.

[26] H. Zhu, X. Zuo, S. Wang, X. Cao, and R. Yang, "Detailed human shape estimation from a single image by hierarchical mesh deformation," in *Proc. CVPR*, 2019.

[27] M. Habermann, W. Xu, M. Zollhoefer, G. Pons-Moll, and C. Theobalt, "A deeper look into DeepCap," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2021.

[28] Y. Feng, V. Choutas, T. Bolkart, D. Tzionas, and M. J. Black, "Collaborative regression of expressive bodies using moderation," in *Proc. 3DV*, 2021.

[29] Z. Zhang, L. Sun, Z. Yang, L. Chen, and Y. Yang, "Global-correlated 3D-decoupling transformer for clothed avatar reconstruction," in *Proc. NeurIPS*, 2023.

[30] X. Yang, Y. Luo, Y. Xiu, W. Wang, H. Xu, and Z. Fan, "D-IF: Uncertainty-aware human digitization via implicit distribution field," in *Proc. CVPR*, 2023.

[31] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *Proc. ECCV*, 2020.

[32] H. Zhang, F. Li, J. Zhao, C. Tan, D. Shen, Y. Liu, and T. Yu, "Controllable free viewpoint video reconstruction based on neural radiance fields and motion graphs," *IEEE Trans. on Visualization and Computer Graphics*, 2022.

[33] S. Hu, F. Hong, L. Pan, H. Mei, L. Yang, and Z. Liu, "SHERF: Generalizable human NeRF from a single image," in *Proc. ICCV*, 2023.

[34] Y. Huang, H. Yi, W. Liu, H. Wang, B. Wu, W. Wang, B. Lin, D. Zhang, and D. Cai, "One-shot implicit animatable avatars with model-based priors," in *Proc. ICCV*, 2023.

[35] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. ICML*, 2021.

[36] V. Gabeur, J.-S. Franco, X. Martin, C. Schmid, and G. Rogez, "Moulding humans: Non-parametric 3D human shape estimation from single images," in *Proc. ICCV*, 2019.

[37] Y. Xian, J. Chibane, B. Lal Bhatnagar, B. Schiele, Z. Akata, and G. Pons-Moll, "Any-shot GIN: Generalizing implicit networks for reconstructing novel classes," in *Proc. 3DV*, 2022.

[38] J. Chibane, T. Alldieck, and G. Pons-Moll, "Implicit functions in feature space for 3D shape reconstruction and completion," in *Proc. CVPR*, 2020.

[39] M. Kazhdan, "Poisson surface reconstruction," in *Proc. SGP*, 2006.

[40] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "DreamFusion: Text-to-3D using 2D diffusion," in *Proc. ICLR*, 2023.

[41] Z. Weng, J. Liu, H. Tan, Z. Xu, Y. Zhou, S. Yeung-Levy, and J. Yang, "Template-free single-view 3D human digitalization with diffusion-guided lrm," *arXiv preprint arXiv:2401.12175*, 2023.

[42] Y. Hong, K. Zhang, J. Gu, S. Bi, Y. Zhou, D. Liu, F. Liu, K. Sunkavalli, T. Bui, and H. Tan, "LRM: Large Reconstruction Model for Single Image to 3D," in *Proc. ICLR*, 2023.

[43] P. Li, W. Zheng, Y. Liu, T. Yu, Y. Li, X. Qi, X. Chi, S. Xia, Y.-P. Cao, W. Xue *et al.*, "PSHuman: Photorealistic single-image 3D human reconstruction using cross-scale multiview diffusion and explicit remeshing," in *Proc. CVPR*, 2025.

[44] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2019.

[45] W. Lorensen and H. Cline, "Marching Cubes: A high resolution 3D surface construction algorithm," *ACM Trans. Graphics*, 1987.

[46] Q. Ma, J. Yang, A. Ranjan, S. Pujades, G. Pons-Moll, S. Tang, and M. J. Black, "Learning to Dress 3D People in Generative Clothing," in *Proc. CVPR*, 2020.

[47] T. Yu, Z. Zheng, K. Guo, P. Liu, Q. Dai, and Y. Liu, "Function4D: Real-time human volumetric capture from very sparse consumer RGBD sensors," in *Proc. CVPR*, 2021.

[48] J. S. Hsuan-I Ho, Lixin Xue and O. Hilliges, "Learning locally editable virtual humans," in *Proc. CVPR*, 2023.

[49] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. CVPR*, 2019.

[50] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Computer Science*, 2014.

[51] T. He, Y. Xu, S. Saito, S. Soatto, and T. Tung, "ARCH++: Animation-ready clothed human reconstruction revisited," in *Proc. ICCV*, 2021.

[52] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[53] S. Lin, L. Yang, I. Saleemi, and S. Sengupta, "Robust high-resolution video matting with temporal guidance," in *Proc. WACV*, 2021.

[54] S. Goel, G. Pavlakos, J. Rajasegaran, A. Kanazawa, and J. Malik, "Humans in 4D: Reconstructing and tracking humans with transformers," in *Proc. ICCV*, 2023.

Qiao Feng (Student Member, IEEE) received his B.Eng and M.Eng degrees in computer science from Tianjin University, Tianjin, China, in 2021 and 2024, respectively. His research interests include 3D Vision, with a current focus on 3D humans.

Kun Li (Senior Member, IEEE) received the B.E. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2006, and the master and Ph.D. degrees from Tsinghua University, Beijing, in 2011. She is currently a Professor with the College of Intelligence and Computing, Tianjin University, Tianjin, China. She was the recipient of the CSIG Shi Qingyun Award for Women Scientists. Her research interests include 3D reconstruction and AIGC.

Yuanwang Yang (Student Member, IEEE) received the B.E. degree from Hebei University of Technology, Tianjin, China, in 2023. He is currently pursuing the M.E. degree in computer science in Tianjin University. His research interests include computer vision, 3D and computer graphics.

Yebin Liu (Member, IEEE) received the B.E. degree from the Beijing University of Posts and Telecommunications, China, in 2002 and the Ph.D. degree from the Department of Automation, Tsinghua University, Beijing, China, in 2009. He is currently a full Professor with Tsinghua University. He was a research fellow in the Computer Graphics Group of the Max Planck Institute for Informatik, Germany, in 2010. His research areas include computer vision, computer graphics, and computational photography.

Yu-Kun Lai (Senior Member, IEEE) received his bachelor's and Ph.D. degrees in computer science from Tsinghua University in 2003 and 2008, respectively. He is currently a professor in the School of Computer Science & Informatics, Cardiff University. His research interests include computer graphics, geometry processing, image processing and computer vision. He is on the editorial boards of IEEE Transactions on Visualization and Computer Graphics and The Visual Computer.

Jingyu Yang (Senior Member, IEEE) received the B.E. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2003, and the Ph.D. degree (Hons.) from Tsinghua University, Beijing, in 2009. He has been a Faculty Member with Tianjin University, Tianjin, China, since 2009, where he is currently a Professor with the School of Electrical and Information Engineering. He served as the Special Session Chair in VCIP 2016 and the Area Chair in ICIP 2017. His research focuses on image processing, 3-D imaging, and computer vision.