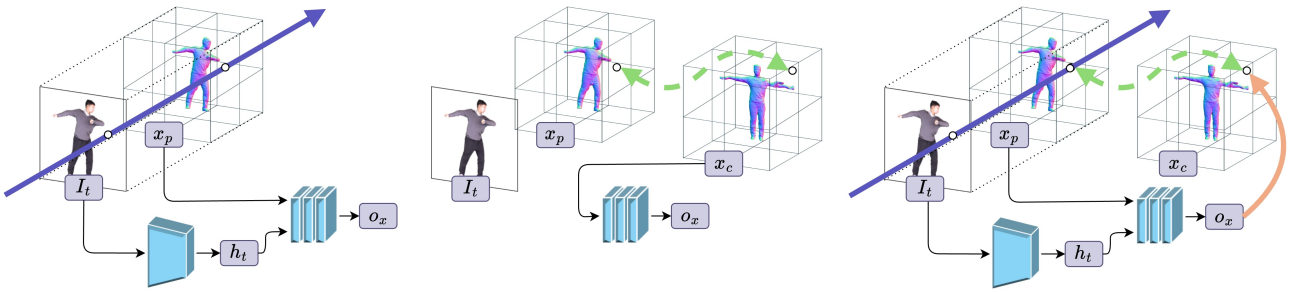
Link to the Past: Temporal Propagation for Fast 3D Human Reconstruction from Monocular Video

Matthew Marchellus, Nadhira Noor, and In Kyu Park
Department of Electrical and Computer Engineering, Inha University
Incheon 22212, Korea
{marchellusmatthew@gmail.com, nadhirannoor@gmail.com, pik@inha.ac.kr}



(a) Pixel-Aligned Reconstruction Methods (b) Video-based Reconstruction Methods (c) TemPoFast3D

Figure 1. We propose **TemPoFast3D**, a novel pipeline to leverage the temporal coherency of human appearance for efficient and accurate 3D human reconstruction from monocular videos. We temporally propagate information from the past frames result by blending the pixel-aligned implicit function and avatar reconstruction method.

Abstract

Fast 3D clothed human reconstruction from monocular video remains a significant challenge in computer vision, particularly in balancing computational efficiency with reconstruction quality. Current approaches are either focused on static image reconstruction but too computationally intensive, or achieve high quality through per-video optimization that requires minutes to hours of processing, making them unsuitable for real-time applications. To this end, we present TemPoFast3D, a novel method that leverages temporal coherency of human appearance to reduce redundant computation while maintaining reconstruction quality. Our approach is a "plug-and play" solution that uniquely transforms pixel-aligned reconstruction networks to handle continuous video streams by maintaining and refining a canonical appearance representation through efficient coordinate mapping. Extensive experiments demonstrate that TemPoFast3D matches or exceeds state-of-the-art methods across standard metrics while providing high-quality textured reconstruction across diverse pose and appearance, with a maximum speed of 12 FPS.

1. Introduction

Real-time 3D human reconstruction from monocular video streams is a fundamental challenge that could revolutionize virtual reality, telepresence, and human-computer interaction. These applications demand methods that can accurately reconstruct 3D clothed humans from single images or video streams, capturing both detailed geometry and realistic appearance. However, existing approaches often struggle to balance the computational efficiency required for video with the high fidelity demanded by real-world applications.

Recent approaches to 3D clothed human reconstruction primarily follow two distinct paradigms. (i) Single-image reconstruction methods rely on pixel-aligned features [1, 9, 30, 31, 36, 37, 43, 45] to capture detailed geometry and textures through implicit functions. (ii) Video-based reconstruction methods leverage pose deformation with either implicit neural fields [7, 13–15, 27, 33–35] or Gaussian splatting [11, 23, 28, 32] for articulated 3D modeling. While producing high-quality results, these methods require extensive optimization (*i.e.* minutes to hours) and multiple passes over the video, making them unsuitable for real-

time applications. Recent attempts at real-time reconstruction [6] either sacrifice quality, require additional inputs (templates [8, 38], multi-view [3, 42], depth [5, 24, 25]), or lack true 3D capability [6, 17, 29], leaving fast, high-quality reconstruction from video an unsolved challenge.

Our key insight is that while human poses change rapidly across video streams, the underlying body shape and clothing geometry remain largely consistent over short time periods. Video-based methods exploit this temporal coherence, but require multiple passes over the entire sequence for global optimization. We observe that this temporal consistency can be leveraged even further for faster reconstruction through progressive canonical shape learning for sequential frame-by-frame processing. However, the challenge of learning a canonical shape through sequential frame-by-frame processing remains largely unexplored in existing literature. We address this limitation by combining the reconstruction accuracy of pixel-aligned methods [1, 9, 30, 31, 36, 37, 43, 45] with bidirectional canonical-posed space mapping [7, 13–15, 27, 33–35], enabling efficient shape learning while maintaining high reconstruction quality from video streams.

To that end, we propose **TemPoFast3D**, a novel fast frame-by-frame 3D reconstruction approach that temporally propagates canonical shape information across video frames. Our method combines pixel-aligned reconstruction with SMPL-based coordinate mapping to maintain a consistent canonical representation while accurately capturing pose variations. This combination naturally extends to multi-view settings when additional views are available. Furthermore, we develop optimization strategies including adaptive coordinate sampling and visibility-guided filtering that significantly reduce per-frame computation. Our framework is designed as a "plug-and-play" solution that accelerates existing SMPL-guided pixel-aligned reconstruction methods, reaching maximum speed of 12 frames per second while maintaining reconstruction quality. In summary, our contributions are:

- We introduce TemPoFast3D, a novel framework combining canonical space inference with efficient pose deformation, enabling faster 3D clothed human reconstruction from monocular video without additional inputs or templates
- A suite of optimization strategies including adaptive coordinate sampling and visibility-guided filtering mechanisms that significantly reduce per-frame computation while preserving reconstruction quality
- A plug-and-play design that accelerates existing SMPL-aligned reconstruction methods while enabling additional capabilities such as multi-view reconstruction for enhanced quality

2. Related Works

Pixel-Aligned Features for Monocular Human Reconstruction. Pixel-aligned reconstruction methods have revolutionized 3D clothed human reconstruction, with PIFu [30] introducing implicit functions that map 2D pixel features to 3D space for enhanced detail representation. Subsequent approaches like PIFuHD [31] and Geo-PIFu [9] further improved reconstruction quality through multi-resolution designs and geometric priors. While effective for visible details, these methods [9, 17, 30, 31] struggle with complex clothing and poses due to their reliance on visible data only. Recent advances integrate parametric models through SMPL [20] framework [2, 10, 12, 18, 45] and improve fidelity using normal maps and SDF prediction [1, 4, 36, 37, 39, 43, 44]. However, these methods prioritize quality over speed, resulting in high computational costs that limit their applicability for video streams where reconstruction speed is crucial.

Pose Deformation for Monocular Video Reconstruction. Leveraging the temporal coherence of human shape from video requires an articulated human model to simulate the natural movement of the human body. The SMPL parametric 3D body model [20] is an articulated human model that contains mesh deformation to adjust the surface mesh (skin and clothing) according to the skeleton's motions, maintaining realistic human contours. Recent works leveraged the deformation capabilities from SMPL model [20] with implicit neural fields (NeRF) [22] to enable high-quality dynamic reconstruction [13–15, 26, 27, 33–35], with Vid2Avatar [7] using pose-conditioned implicit signed-distance fields for geometry and texture representation. Despite producing realistic results across various poses, these approaches require minutes or hours of training with relatively slow rendering speeds. Recent approaches have leveraged Gaussian splatting techniques [11, 16, 19, 23, 28] for improved rendering efficiency while still relying on offline optimization. While these video-based methods produce higher quality results than pixel-aligned approaches, their extensive per-video optimization requirements limit their applicability to pre-recorded videos rather than real-time applications.

Real-time 3D clothed human reconstruction. Existing approaches to real-time 3D clothed human reconstruction face significant limitations. Methods like Monoport [17] achieve near real-time performance by bypassing explicit 3D reconstruction, while [29] offers real-time rendering but requires hours for the actual reconstruction. Multi-view approaches [3, 42] require calibrated camera setups unsuitable for in-the-wild scenarios, and template-based methods [8, 38] depend on pre-scanned templates that limit generalization. Depth-based techniques [5, 24, 25] rely on specialized sensors, while FOF [6] achieves 30 FPS but lacks

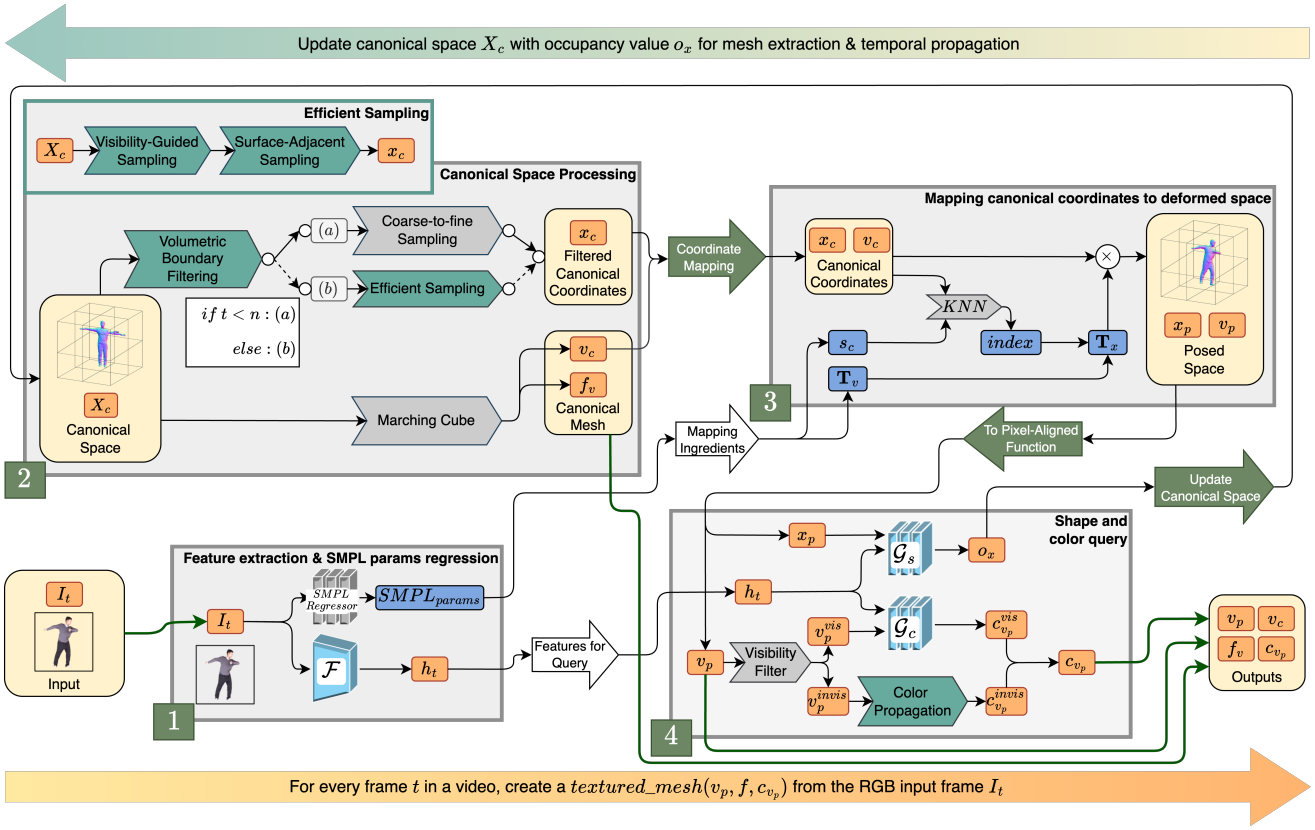


Figure 2. **Overview of our TemPoFast3D pipeline.** Given an input RGB frame I_t, our method combines efficient canonical space processing with coordinate mapping for fast 3D human reconstruction. The pipeline consists of: (Section 3.1) Feature extraction and SMPL params regression, (Section 3.2) Mapping canonical coordinates to posed space, (Section 3.2.2) Shape and color query, and (Section 3.3) Canonical space processing. The canonical space representation X_c is continuously updated across frames.

texture inference capability. These limitations underscore the need for methods that balance speed and quality without requiring additional hardware or subject-specific templates.

3. Proposed Methods

We present **TemPoFast3D**, a novel plug-and play pipeline for fast 3D clothed human reconstruction from monocular video that combines pixel-aligned features with coordinate mapping in canonical space (See Figure 2). Our method introduces temporal propagation strategies through volumetric boundary filtering and visibility-guided sampling to achieve faster performance while maintaining reconstruction quality. The pipeline's "plug-and-play" design allows the feature extraction network $\mathcal{F}$ and query networks $\mathcal{G}_s, \mathcal{G}_c$ to be replaced with any SMPL-guided pixel-aligned backbone, enabling easy integration of future improvements. We first discuss preliminaries in Section 3.1, then detail our canonical space inference framework in Section 3.2, and finally present our temporal propagation strategy in Section 3.3. As a side benefit from connecting pixel-aligned reconstruction to canonical space, our method can naturally extend to reconstruct multi-view data without any modification will be explained in Section 3.4.

3.1. Preliminary

Pixel-aligned Implicit Function Pixel-aligned Implicit Function (PIFu) [30] enable volumetric reconstruction from a single image by learning a mapping between 2D pixel features and 3D occupancy. Given an input image $\mathbf{I}$, the method first processes it through a filter network $\mathcal{F}$ to obtain a feature map $h = \mathcal{F}(\mathbf{I})$ that captures spatial and visual information. For reconstruction, the method samples 3D query points $\mathbf{x} \in \mathbb{R}^3$ in the camera frustum and projects them onto the image plane to obtain relative pixel coordinates $\mathbf{p}_x = (x, y)$ normalized to $[-1, 1]$. A shape query network $\mathcal{G}_s$ then predicts occupancy values by combining the projected coordinates with pixel-aligned features:

$$\mathbf{o} = \mathcal{G}_s(h, \mathbf{x}_p), \begin{cases} 1 & \text{if } \mathbf{x}_p \text{ lies inside the surface} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Similarly, a color query network $\mathcal{G}_c$ predicts RGB values $\mathbf{c} \in \mathbb{R}^3$ at query points using the same pixel-aligned fea-

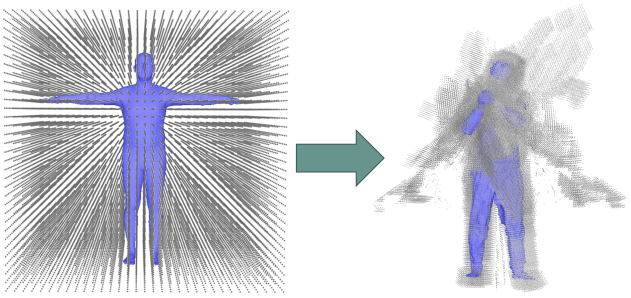


Figure 3. **Warped sampling coordinate visualization.** Sampling points (gray) transition from uniform distribution in canonical space (left) to non-uniform distribution after deformation (right), demonstrating how our coordinate mapping affects sampling density around the SMPL mesh (blue).

tures:

$$\mathbf{c} = \mathcal{G}_c(h, \mathbf{x}_p) \tag{2}$$

Canonical-Posed Space Transformation. To establish a mapping between canonical and posed spaces, we leverage the transformation mechanism derived from SMPL [20]. The SMPL model defines a parametric function $\mathcal{S}(\boldsymbol{\beta}, \boldsymbol{\theta})$, where $\boldsymbol{\beta} \in \mathbb{R}^{10}$ represents shape parameters and $\boldsymbol{\theta} \in \mathbb{R}^{3\times 24}$ defines pose parameters. The transformation between spaces is computed through Linear Blend Skinning:

$$\mathbf{T}_s = \sum_{k=1}^{K} w_{k,i} \mathbf{G}_k(\boldsymbol{\beta}, \boldsymbol{\theta}) \tag{3}$$

where $w_{k,i}$ are blend weights and $\mathbf{G}_k(\boldsymbol{\beta}, \boldsymbol{\theta})$ are the joint transformation matrices computed from SMPL parameters. The resulting vertices transformation matrices $\mathbf{T}_s \in \mathbb{R}^{N\times 4\times 4}$ enable bidirectional mapping of SMPL vertices:

$$\text{Forward mapping:} \quad \mathbf{s}_p = \mathbf{T}_s \mathbf{s}_c \tag{4}$$

$$\text{Inverse mapping:} \quad \mathbf{s}_c = \mathbf{T}_s^{-1} \mathbf{s}_p \tag{5}$$

These per-vertex transformation matrices are computed once per frame through the SMPL layer and can be transferred to arbitrary coordinates through association with the average of nearest SMPL vertices.

3.2. Canonical Space Inference and Coordinate Mapping

Leveraging pixel-aligned reconstruction methods while facilitating temporal information sharing necessitates performing shape inference in canonical space rather than posed space. This requires establishing a bidirectional mapping between canonical coordinates and posed space where pixel-aligned features are computed. Given a set of canonical coordinates $\mathbf{x}_c$, we establish correspondence with SMPL canonical vertices $\mathbf{s}_c$ through K[1] nearest neighbor search. Using this correspondence, we transfer the SMPL vertex transformations $\mathbf{T}_s$ to create coordinate transformations $\mathbf{T}_x \in \mathbb{R}^{N\times 4\times 4}$, effectively extending the SMPL deformation field to arbitrary points in space. Following Eq. 4, the transformation to is then computed as:

$$\mathbf{x}_p = \mathbf{T}_x \mathbf{x}_c \tag{6}$$

Note that the deformation process indiscriminately maps all canonical space coordinates - both inside and outside the body shape - into regions around the deformed body, where the geometric consistency of the reconstruction can be affected by external canonical points being mapped into the interior of the deformed body as illustrated in Figure 3.

3.2.1. Volumetric Boundary Filtering

We observe that valid canonical human geometry predominantly resides within a proximal volume around the canonical SMPL mesh $\mathbf{s}_c$. A simple discrete volumetric boundary as a mask to this proximal region in canonical space is sufficient to eliminate irrelevant query points. We formulate this mask as a binary spatial classifier where

$$\mathbf{m}_i = \begin{cases} 1 & \text{if } \mathbf{x}_c^i \text{ lies within boundary} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

As shown in Figure 4, this filtering mechanism prevents the inclusion of irrelevant canonical coordinates in the shape inference process while additionally reducing the total number of query points required for reconstruction.

3.2.2. Posed Mesh Generation and Color Inference

From the canonical occupancy field, we initially extract an isosurface mesh in canonical pose using the Marching Cubes algorithm. To obtain the posed configuration, we apply a similar deformation procedure as defined in Eq. 6 to transform the mesh into posed space. Specifically, given canonical vertices $\mathbf{v}_c$, we establish vertex-specific transformation matrices $\mathbf{T}_m$ through nearest neighbor correspondence with canonical SMPL vertices $\mathbf{s}_c$ using K nearest neighbor search, following our earlier coordinate mapping formulation. The final posed vertices $\mathbf{V}_p$ are then computed by applying these transformations:

$$\mathbf{v}_p = s(\mathbf{T}_m \mathbf{v}_c + \mathbf{t}) \tag{8}$$

where s and $\mathbf{t}$ are the scale and translation parameters from SMPL estimation. This transformation enables us to leverage pixel-aligned features for color prediction, as the posed vertices $\mathbf{v}_p$ now align with the input image's space. Using these pixel-aligned vertices, we employ the color query function G_c from Section 3.1 to predict color values directly from the input image features. The bidirectional mapping

[1]Details on the number of K neighbor is provided in supplementary.

between canonical and posed space enables seamless integration with existing pixel-aligned reconstruction methods for both shape and color inference, while extending the applicability of these methods to temporal sequences.

3.3. Temporal Propagation and Efficient Inference

While single-frame reconstruction follows the coordinate mapping described in Section 3.2, video sequences offer opportunities for additional computational optimization by leveraging temporal coherency. We introduce a frame threshold n that determines when to transition from full reconstruction to efficient inference, as illustrated at the top left of Figure 2. Let t denote the current frame index:

$$\text{Sampling Strategy} = \begin{cases} \text{Coarse-to-fine sampling} & \text{if } t \leq n, \\ \text{Efficient inference} & \text{if } t > n \end{cases} \tag{9}$$

For the initial n frames, we perform complete shape inference to establish a reliable canonical shape representation. After frame n, we transition to our efficient inference strategy that enables four key optimizations:

3.3.1. Bypass Coarse-to-Fine Inference

Temporal propagation strategy establishes a robust geometric prior through propagated canonical shape from previous frames $\mathbf{v}_c^{prev}$, rendering hierarchical coarse-to-fine inference redundant. This bypass is particularly advantageous in our framework, as each query point incurs additional computational overhead from coordinate mapping operations. Instead of predicting the entire volume, we focus computation exclusively on regions requiring refinement, reducing the per-frame query complexity. To identify these regions efficiently, we employ two complementary sampling strategies based on visibility (Section 3.3.2) and surface proximity (Section 3.3.3).

3.3.2. Visibility-Guided Sampling

Given a prior canonical shape $\mathbf{v}_c^{prev}$, point queries need only be performed on coordinates that are observable from the current viewpoint, as these regions yield the most reliable predictions from pixel-aligned features. To identify these regions, we first compute a visibility mask $\mathbf{m}_v$ for the canonical SMPL vertices $\mathbf{s}_c$ through mesh rasterization. Following our established attribute transfer mechanism (Section 3.2), the visibility status is then propagated to canonical coordinates $\mathbf{x}_c$ through K nearest neighbor search and thus filters out coordinates from unseen region.

3.3.3. Surface-Adjacent Sampling

To achieve better computational efficiency, we constrain point queries to regions near the surface boundary. Specifically, we sample points where the occupancy value o_c falls within a narrow band defined by thresholds α and β:

$$\alpha \leq o_c \leq \beta \tag{10}$$

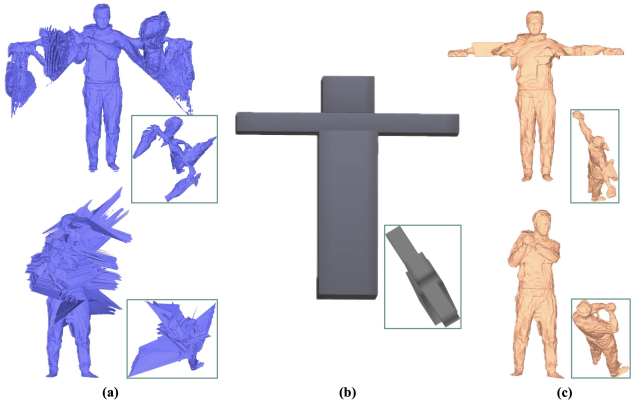


Figure 4. **Effect of volumetric boundary filtering.** (a) Reconstructed meshes without filtering in canonical (top) and deformed pose (bottom) show artifacts. (b) Volumetric boundary mask. (c) Filtered reconstruction results show cleaner geometry in both poses, eliminating artifacts beyond the valid body region.

This targeted sampling strategy enables us to maintain detailed surface geometry while significantly reducing query points required for inference.

3.3.4. Color Propagation and Visibility Handling

For texture inference, we introduce a visibility-aware color propagation strategy that leverages the canonical space representation to handle occluded regions effectively. Given a deformed mesh with vertices $\mathbf{v}_d$, we first predict colors for visible vertices $c_{v_p}^{vis}$ using the color query network $\mathcal{G}_c$. For vertices that are occluded or poorly visible in the current frame, we employ a neighbor-based color propagation scheme that operates in canonical space. Specifically, we establish correspondence between current canonical vertices $\mathbf{v}_c$ and previous canonical vertices $\mathbf{v}_c^{prev}$ through K nearest neighbor search, thus obtaining color $c_{v_p}^{invis}$ from correspondence. We obtain the final color c_{v_p} at time by combining both $c_{v_p}^{vis}$ and $c_{v_p}^{invis}$.

3.4. Extension to Multi-View Inference

The coordinate mapping strategy and canonical inference mechanism (Section 3.2) naturally extend to multi-view reconstruction scenarios, allowing us to aggregate information from synchronized viewpoints without architectural modifications. By independently processing each view through our pipeline and merging their canonical representations, we leverage the unified canonical space as a consistent global reference for both temporal and spatial fusion. While not specifically optimized for multi-view scenarios, this capability provides additional validation of our framework's ability to accumulate and refine geometric details through multiple observations.

Method	CAPE-NFP			CAPE-FP			THuman2.0			
	Chamfer ↓	P2S ↓	Normal ↓	Chamfer ↓	P2S ↓	Normal ↓	Chamfer ↓	P2S ↓	Normal ↓	PSNR ↑
PIFu∗ [30]	2.5609	1.9971	0.1023	1.8139	1.5108	0.0798	1.5991	1.4333	0.0843	18.09
PIFuHD∗ [31]	3.7670	3.5910	0.1230	2.3020	2.3350	0.0900	-	-	-	-
ECON∗ [37]	0.9462	0.9334	0.0382	0.9039	0.8938	0.0373	1.2585	1.4184	0.0612	-
GTA∗ [43]	0.8508	0.7920	0.0424	0.6525	0.6084	0.0349	0.7329	0.7297	0.0492	18.05
SIFU∗ [44]	**0.7725**	0.7354	0.0378	0.6297	0.5980	0.0327	0.5961	0.6058	0.0407	22.10
PIFu† [30]	4.2310	4.7087	0.1029	2.5917	2.8163	0.0827	3.1788	3.3589	0.1082	-
GTA† [43]	0.9160	0.8482	0.0429	0.6531	0.6084	0.0347	0.4625	0.4677	0.0348	**23.27**
SIFU† [44]	0.8263	0.7889	0.0384	0.6254	0.5901	0.0323	0.4409	0.4580	0.0342	22.82
TPF3D-GTA	0.9939	0.7724	0.0507	0.7057	0.5841	0.0383	0.5247	0.4530	0.0383	23.25
TPF3D-SIFU	0.9230	0.7147	0.0464	0.6833	0.5663	0.0359	0.5047	0.4432	0.0374	22.69
TPF3D-GTA-3v	0.8293	0.6587	0.0391	0.5855	0.4967	0.0285	0.4195	0.3632	**0.0307**	23.21
TPF3D-SIFU-3v	0.8024	**0.6351**	**0.0370**	**0.5794**	**0.4883**	**0.0278**	**0.4144**	**0.3590**	0.0313	22.66

Table 1. **Quantitative comparison against state-of-the-art methods.** ∗: Results of the compared methods obtained from [44], †: We re-evaluated the compared methods for a fair comparison in the same environment (*cf.* Section 4 and Figure 5).

4. Experiment

Implementation Details. Our pipeline is implemented in PyTorch and executed on a single NVIDIA RTX 4090 GPU. We utilize PyMAF [41] for SMPL [20] parameter regression on in-the-wild data. We evaluate our pipeline with GTA [43] and SIFU [44] as our pixel-aligned networks (denoted as TPF3D-GTA and TPF3D-SIFU respectively), utilizing their original pre-trained weights to demonstrate our method's plug-and-play capability. Following the multi-view extension described in Section 3.4, we also evaluate three-view configurations (TPF3D-GTA-3v, TPF3D-SIFU-3v) where orthogonal views (0°, 120°, 240°) are used to enhance reconstruction quality. The canonical space reconstruction operates at 256^3 resolution. For temporal propagation, we set the frame threshold $n^2 = 5$ before transitioning to efficient inference (see supplementary for detailed analysis). Surface-adjacent sampling uses thresholds $\alpha = 0.4$ and $\beta = 0.7$ to define the sampling region, with random shuffling applied after thresholding for better coverage.

Datasets. Our pre-trained weight for the pixel-aligned reconstruction networks are trained exclusively on the THuman2.0 dataset [40], which consists of 526 human scans along with their corresponding SMPL-X fits. Of these, 490 are allocated for training, 15 for validation, and 21 for testing. For zero-shot evaluation, we use the CAPE dataset [21]. Following previous works, we divide the CAPE dataset into "CAPE-FP" and "CAPE-NFP" instead of "fashion" and "non-fashion" poses, respectively. Video performance is evaluated on the NeuMan dataset [15], using the *bike, citron, jogging, and seattle* sequences as per their official testing splits, following previous works [11, 23].

Evaluation Metrics. We employ chamfer distance and P2S (point-to-surface) to evaluate geometric error between ground-truth and predicted mesh. For reconstruction on single images, shape surface detail and consistency is evaluated using L2 normal error, while texture is evaluated with PSNR. We utilize a combination of PSNR, SSIM, and LPIPS to evaluate reconstruction accuracy on video data following [23]. Quality and speed trade-offs are emphasized by comparing the average FPS for inference/rendering and training time for each video-based method.

[2]Details on hyperparameter n is provided in supplementary.

4.1. Evaluation

Evaluation on THuman2.0 [40] and CAPE [21]. We first evaluate its performance on single-image datasets to establish baseline capabilities. Table 1 shows quantitative comparisons with SOTA single-image reconstruction methods on THuman2.0 [40] and CAPE [21] datasets. It should be noted that fair comparison is not possible as TPF3D requires multiple view/frame for optimal result while other methods only need single image to achieve max quality. That being said, Table 1 shows that TPF3D slightly degrades the quality of single-view result but achieves higher reconstruction quality after merging multiple frames (*i.e.*, **TPF3D-SIFU-3v** and **TPF3D-GTA-3v**). These scores demonstrate that our approach effectively leverages temporal coherency without sacrificing reconstruction quality. Figure 5 presents qualitative comparisons across various poses and clothing styles. Our method successfully captures fine geometric details such as clothing wrinkles and body contours, particularly in challenging cases like raised arms (row 3) and twisted poses (row 4). While single-frame methods like GTA [43] and SIFU [44] achieve impressive results considering their limited input, our method's ability to combine multiple-view produces a smoother result at the cost of some artifacts due to pose deformation. For texture reconstruction, our approach achieves comparable PSNR (23.25 dB) to the re-evaluated GTA (23.27 dB). Additional comparisons are provided in the supplementary material.

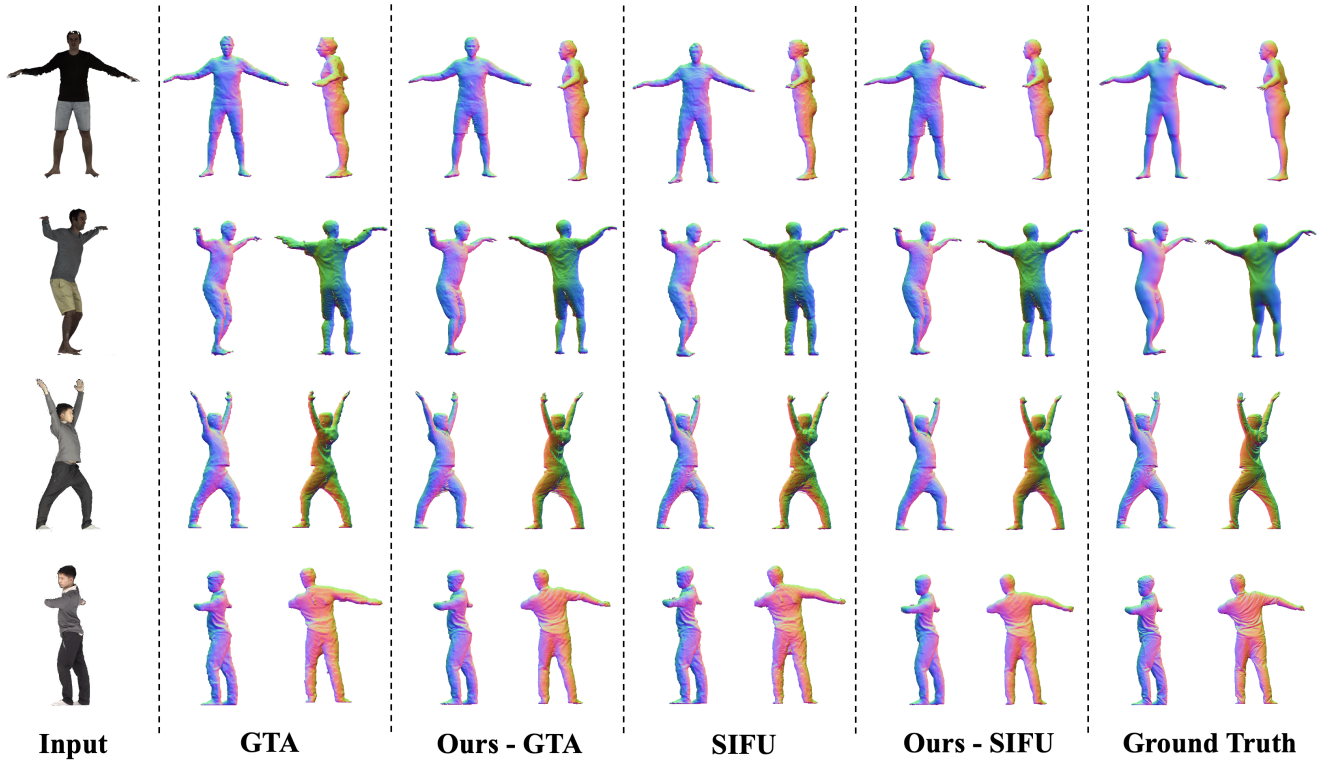


Figure 5. **Qualitative comparison of geometry reconstruction quality.** The top two rows show results on the CAPE dataset [21], while the bottom two rows are from the THuman2.0 dataset [40]. For best viewing, please zoom in on a digital screen.

Evaluation on NeuMan [15]. We simulate zero-shot inference using in-the-wild videos from NeuMan dataset without their ground truth SMPL parameters. Similar to evaluation on single-frame reconstruction, fair comparison is not possible as our method does not require per-subject optimization contrary to other methods. Table 2 shows that our method (TPF3D-GTA) are comparable against early optimization-based approaches in terms of rendering quality (i.e., HumanNerf [35]). ExAvatar [23] achieves much higher accuracy at the cost of 4 hours of training and slow rendering speed. GaussianAvatar [11] and InstantAvatar [14] achieves faster rendering speed compared to our method, though it should be noted that they require per-subject optimization while our lower FPS includes shape and color reconstruction. Overall, Table 2 highlights the reconstruction quality and fast speed trade-off for video reconstruction methods, demonstrating that our approach provides a balanced solution with competitive performance and efficient inference. For qualitative comparison, we compare our reconstruction result with the ground truth in Figure 6 for frame 2, 12, and 22 from the *bike* sequences. TPF3D-GTA method achieves high-quality geometry and texture reconstruction from as early as frame 2, with facial details becoming increasingly clear in later frames. Minor texture inconsistencies are attributable to our vertex-based color representation.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	Training ↓	Avg. FPS ↑
HumanNerf [35]	27.06	0.967	0.019	26h 33m	0.540
InstantAvatar [14]	28.47	0.972	0.028	00h 26m	**21.000**
NeuMan [15]	29.32	0.958	0.014	128h 00m∗	0.004
Vid2Avatar [7]	30.70	0.980	0.014	97h 01m	0.008
GaussianAvatar [11]	29.94	0.980	0.012	00h 43m	15.720
3DGS-Avatar [28]	28.99	0.974	0.016	-	-
ExAvatar [23]	**34.80**	**0.984**	**0.009**	04h 0m	4.022
TPF3D-GTA (Ours)	27.60	0.965	0.022	**pretrained**	8.900

Table 2. **Quantitative evaluation on NeuMan [15].** PSNR, SSIM, and LPIPS results of other methods are taken from [23]. We run each methods in the same environment[3] for a fair speed comparison, except "∗" is obtained from [15] (*cf.* Section 4.1).

4.2. Ablation Studies

We evaluate our optimization strategies on the *citron* sequence from the NeuMan dataset, as detailed in Table 3. Our baseline implementation achieves 3.27 FPS with PSNR of 32.80 using GTA [43]. Coordinate mapping initially reduces speed, but enables optimizations such as surface-adjacent sampling for a significant speed up, and improved further with limiting the sampling points. TorchScript optimization provides the final breakthrough, achieving a maximum of 12.3 FPS - a 3.06× speedup over baseline. Throughout these optimizations, reconstruction quality re-

[3] Additional details are provided in supplementary

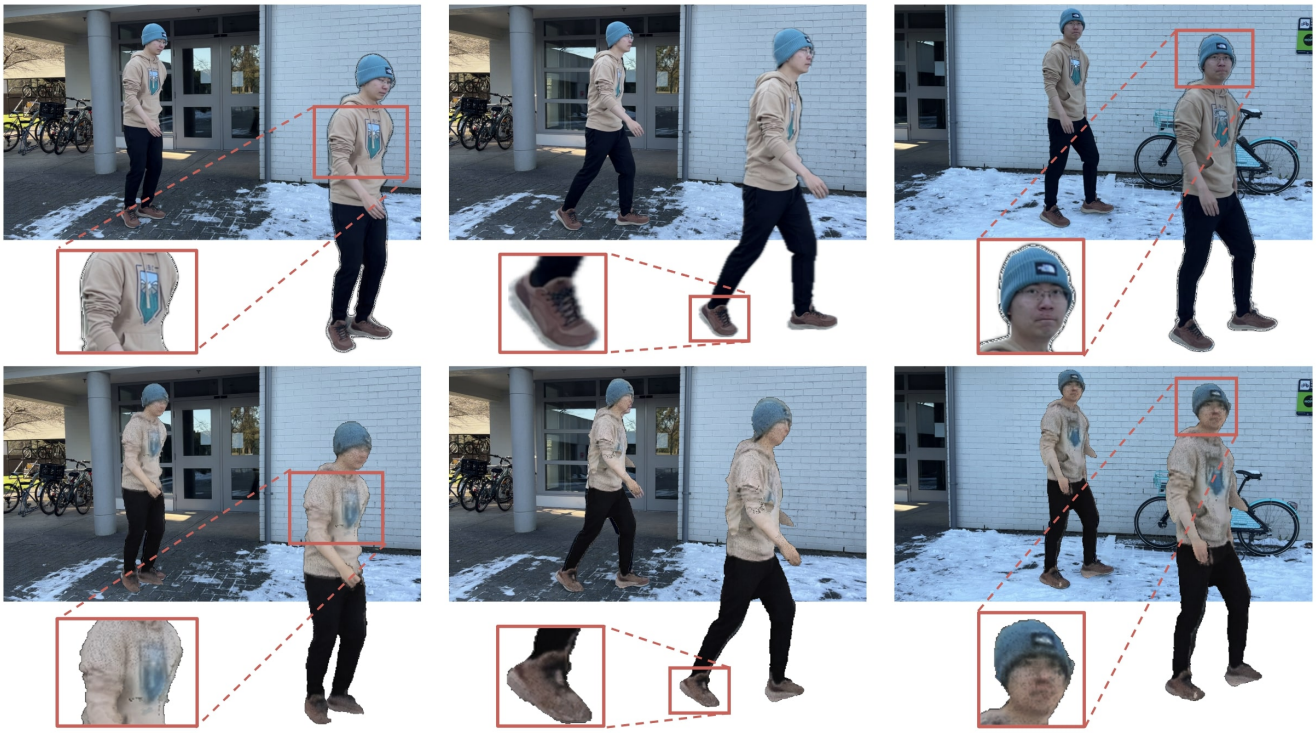


Figure 6. **Texture quality results on NeuMan [15] dataset.** Comparison between ground truth (top) and our real-time reconstruction (bottom) showing consistent quality across early (frame 2) to later frames (frame 22).

Method	Max. FPS ↑	PSNR ↑	SSIM ↑	LPIPS ↓
Base (GTA [43])	3.266	32.800	0.985	0.0090
+ Coordinate mapping	2.142	31.024	0.983	0.0102
+ Linear layer	2.669	30.964	0.983	0.0102
+ Visibility-guided sampling	1.908	30.965	0.983	0.0103
+ Surface-adjacent sampling	4.499	30.961	0.983	0.0102
+ Limit sampling points	5.841	30.998	0.983	0.0102
+ Torchscript	12.301	31.132	0.982	0.0106

Table 3. **Ablation study on optimization strategies.** Quantitative comparison of speed (FPS) and quality metrics on NeuMan [15].

mains remarkably stable with PSNR above 30.96, SSIM above 0.98, and LPIPS below 0.011, demonstrating that our speed improvements preserve visual fidelity.

5. Conclusion

We proposed the TemPoFast3D, a novel approach for fast sequential 3D clothed human reconstruction from monocular RGB video stream. We designed the pipeline based on the key idea that human appearance remains largely consistent across video frames, thus complete shape reconstruction every frame was deemed redundant. TemPoFast3D employed canonical space inference using coordinate mapping to establish a canonical space that can be effectively propagated through time. This propagated canonical space allowed our pipeline to maintain and refine shape predictions selectively, reducing the computation cost compared to traditional approaches. Experimental results showed that TemPoFast3D achieves competitive accuracy with state-of-the-art methods while achieving a maximum speed of 12 FPS, pioneering a new direction in efficient 3D human reconstruction through effective temporal information utilization. Despite so, TemPoFast3D exhibit limitations such as occasional artifacts due to pose deformation as sharp protrusions in select few areas. Moreover, our method requires accurate alignment between the predicted mesh and the SMPL model, limiting reconstruction accuracy for loose clothing and accessories. As a result the reconstruction quality depends on the accuracy of SMPL parameter estimation, which are particularly challenging for extreme poses and occlusions.

Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2022-00155915, Artificial Intelligence Convergence Innovation Human Resources Development (Inha University) and No.2021-0-02068, Artificial Intelligence Innovation Hub and IITP-2024-RS-2024-00360227, Leading Generative AI Human Resources Development). This work was supported by Inha University Research Grant.

References

[1] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3D reconstruction of humans wearing clothing. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1496–1505, 2022. 1, 2

[2] Yukang Cao, Kai Han, and Kwan-Yee K. Wong. SeSDF: Self-evolved signed distance field for implicit 3D clothed human reconstruction. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[3] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam G. Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Trans. on Graphics*, 34(4):69:1–69:13, 2015. 2

[4] Enric Corona, Mihai Zanfir, Thiemo Alldieck, Eduard Gabriel Bazavan, Andrei Zanfir, and Cristian Sminchisescu. Structured 3D features for reconstructing controllable avatars. pages 16954–16964. IEEE, 2023. 2

[5] Zijian Dong, Chen Guo, Jie Song, Xu Chen, Andreas Geiger, and Otmar Hilliges. PINA: learning a personalized implicit neural avatar from a single RGB-D video sequence. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[6] Qiao Feng, Yebin Liu, Yu-Kun Lai, Jingyu Yang, and Kun Li. FOF: learning fourier occupancy field for monocular real-time human reconstruction. In *Proc. Advances in Neural Information Processing Systems*, 2022. 2

[7] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2Avatar: 3D avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12858–12868, 2023. 1, 2, 7

[8] Marc Habermann, Weipeng Xu, Michael Zollhöfer, Gerard Pons-Moll, and Christian Theobalt. LiveCap: Real-time human performance capture from monocular video. *ACM Trans. on Graphics*, 38(2):14:1–14:17, 2019. 2

[9] Tong He, John P. Collomosse, Hailin Jin, and Stefano Soatto. Geo-PIFu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In *Proc. Advances in Neural Information Processing Systems*, 2020. 1, 2

[10] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. ARCH++: animation-ready clothed human reconstruction revisited. In *Proc. IEEE/CVF International Conference on Computer Vision*, pages 11026–11036, 2021. 2

[11] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. GaussianAvatar: Towards realistic human avatar modeling from a single video via animatable 3D gaussians. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 634–644, 2024. 1, 2, 6, 7

[12] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. ARCH: animatable reconstruction of clothed humans. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3090–3099, 2020. 2

[13] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. SelfRecon: Self reconstruction your digital avatar from monocular video. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5595–5605, 2022. 1, 2

[14] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. InstantAvatar: Learning avatars from monocular video in 60 seconds. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16922–16932, 2023. 7

[15] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. NeuMan: Neural human radiance field from a single video. In *Proc. European Conference on Computer Vision*, pages 402–418, 2022. 1, 2, 6, 7, 8

[16] Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. HUGS: human gaussian splats. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 505–515, 2024. 2

[17] Ruilong Li, Yuliang Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li. Monocular real-time volumetric performance capture. pages 49–67, 2020. 2

[18] 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 *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8662–8672, 2023. 2

[19] Xinqi Liu, Chenming Wu, Jialun Liu, Xing Liu, Chen Zhao, Haocheng Feng, Errui Ding, and Jingdong Wang. GVA: Reconstructing vivid 3D gaussian avatars from monocular videos. *Arxiv*, 2024. 2

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

[21] 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 *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 6, 7

[22] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Proc. European Conference on Computer Vision*, pages 405–421, 2020. 2

[23] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3D gaussian avatar. In *Proc. European Conference on Computer Vision*, 2024. 1, 2, 6, 7

[24] Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew W. Fitzgibbon. KinectFusion: Real-time dense surface mapping and tracking. In *Proc. IEEE International Symposium on Mixed and Augmented Reality*, pages 127–136, 2011. 2

[25] Richard A. Newcombe, Dieter Fox, and Steven M. Seitz. DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 343–352, 2015. 2

[26] Xiao Pan, Zongxin Yang, Jianxin Ma, Chang Zhou, and Yi Yang. TransHuman: A transformer-based human representation for generalizable neural human rendering. In *Proc.*

IEEE/CVF International Conference on Computer Vision, pages 3521–3532, 2023. 2

[27] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural Body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 1, 2

[28] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3DGS-Avatar: Animatable avatars via deformable 3D gaussian splatting. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5020–5030, 2024. 1, 2, 7

[29] Ignacio Rocco, Iurii Makarov, Filippos Kokkinos, David Novotný, Benjamin Graham, Natalia Neverova, and Andrea Vedaldi. Real-time volumetric rendering of dynamic humans. *arXiv:2303.11898*, 2023. 2

[30] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Hao Li, and Angjoo Kanazawa. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proc. IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019. 1, 2, 3, 6

[31] Shunsuke Saito, Tomas Simon, Jason M. Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 81–90, 2020. 1, 2, 6

[32] Kaiyue Shen, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Julien Valentin, Jie Song, and Otmar Hilliges. X-Avatar: Expressive human avatars. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16911–16921, 2023. 1

[33] Shih-Yang Su, Timur M. Bagautdinov, and Helge Rhodin. DANBO: disentangled articulated neural body representations via graph neural networks. In *Proc. European Conference on Computer Vision*, pages 107–124, 2022. 1, 2

[34] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-NeRF: Articulated neural radiance fields for learning human shape, appearance, and pose. In *Proc. Advances in Neural Information Processing Systems*, 2021.

[35] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16189–16199, 2022. 1, 2, 7

[36] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: implicit clothed humans obtained from normals. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13286–13296, 2022. 1, 2

[37] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: explicit clothed humans optimized via normal integration. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 512–523, 2023. 1, 2, 6

[38] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. MonoPerfCap: Human performance capture from monocular video. *ACM Trans. on Graphics*, 37(2), 2018. 2

[39] Xueting Yang, Yihao Luo, Yuliang Xiu, Wei Wang, Hao Xu, and Zhaoxin Fan. D-IF: uncertainty-aware human digitization via implicit distribution field. In *Proc. IEEE/CVF International Conference on Computer Vision*, pages 9088–9098, 2023. 2

[40] 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 *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 6, 7, 2, 3, 5

[41] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proc. IEEE/CVF International Conference on Computer Vision*, pages 11426–11436, 2021. 6

[42] Jiakai Zhang, Xinhang Liu, Xinyi Ye, Fuqiang Zhao, Yanshun Zhang, Minye Wu, Yingliang Zhang, Lan Xu, and Jingyi Yu. Editable free-viewpoint video using a layered neural representation. *ACM Trans. on Graphics*, 40(4):149:1–149:18, 2021. 2

[43] Zechuan Zhang, Li Sun, Zongxin Yang, Ling Chen, and Yi Yang. Global-correlated 3D-decoupling transformer for clothed avatar reconstruction. In *Proc. Advances in Neural Information Processing Systems*, 2023. 1, 2, 6, 7, 8

[44] Zechuan Zhang, Zongxin Yang, and Yi Yang. SIFU: side-view conditioned implicit function for real-world usable clothed human reconstruction. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9936–9947, 2024. 2, 6

[45] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. PaMIR: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 44(6):3170–3184, 2022. 1, 2