DLCA-Recon: Dynamic Loose Clothing Avatar Reconstruction from Monocular Videos

Chunjie Luo[1], Fei Luo[1] *, Yuseng Wang[1], Enxu Zhao[1], Chunxia Xiao[1]*

[1]School of Computer Science, Wuhan University, Wuhan, China
luochunjie@whu.edu.cn, luofei@whu.edu.cn, wangyusen@whu.edu.cn, zhaoenxu@whu.edu.cn, cxxiao@whu.edu.cn

Abstract

Reconstructing a dynamic human with loose clothing is an important but difficult task. To address this challenge, we propose a method named DLCA-Recon to create human avatars from monocular videos. The distance from loose clothing to the underlying body rapidly changes in every frame when the human freely moves and acts. Previous methods lack effective geometric initialization and constraints for guiding the optimization of deformation to explain this dramatic change, resulting in the discontinuous and incomplete reconstruction surface. To model the deformation more accurately, we propose to initialize an estimated 3D clothed human in the canonical space, as it is easier for deformation fields to learn from the clothed human than from SMPL. With both representations of explicit mesh and implicit SDF, we utilize the physical connection information between consecutive frames and propose a dynamic deformation field (DDF) to optimize deformation fields. DDF accounts for contributive forces on loose clothing to enhance the interpretability of deformations and effectively capture the free movement of loose clothing. Moreover, we propagate SMPL skinning weights to each individual and refine pose and skinning weights during the optimization to improve skinning transformation. Based on more reasonable initialization and DDF, we can simulate real-world physics more accurately. Extensive experiments on public and our own datasets validate that our method can produce superior results for humans with loose clothing compared to the SOTA methods.

Introduction

Reconstructing full-body 3D human models is an important research topic in computer graphics. It has many applications in AR/VR (Bao et al. 2022; Cao et al. 2022, 2023), virtual try-on, and video game industry. Traditionally, high-fidelity human reconstruction requires multi-camera systems, controlled studios, and long-term works of talented artists, making it expensive and highly specialized. Along with the emergence of new applications like digit human in the Metaverse, it demands lightweight and convenient reconstruction solutions to create 3D digital avatars for complex human motions and diverse manners of dressing.

Figure 1: A dynamic loose clothing avatar created by our DLCA-Recon method from a monocular video.

Compared to close-fitting wear, it is more difficult to reconstruct dynamic humans with loose clothing, due to the high freedom of body motions, the appearance diversity, and the deformation randomness of loose clothes. Traditional methods based on explicit mesh are restricted by fixed topologies and resolutions (Alldieck et al. 2018; Guo et al. 2021). Recent methods based on the implicit neural representation for monocular human reconstruction have achieved compelling results (Saito et al. 2019; Huang et al. 2020; He et al. 2020; Gropp et al. 2020; Zheng et al. 2021b; Xiu et al. 2022). These methods can handle arbitrary topologies, enabling the representation of various clothing. However, they require high-quality 3D supervision data and NeRF-based methods usually produce noised geometry.

In addition, there are methods working in a frame-by-frame manner, but they fail to recover invisible parts. Directly regressing 3D surfaces from images is an alternative way (He et al. 2020; Huang et al. 2020; He et al. 2021; Peng et al. 2021a,b; Weng et al. 2022; Li, Luo, and Xiao 2023; Luo et al. 2023). They struggle with out-of-distribution of poses and shapes and cannot carry out temporally continuous 3D reconstructions. SelfRecon (Jiang et al. 2022a) combines explicit and implicit representation to reconstruct a temporally consistent 3D clothed human from a video. But it is restricted to tight clothing and self-rotation movement.

To tackle creating a temporal-spatial coherent 3D model for a human with loose clothing and free movement, we propose a new method DLCA-Recon based on SelfRecon (Jiang

*Fei Luo and Chunxia Xiao are co-corresponding authors.

et al. 2022a). When dealing with the diverse topologies of loose clothing, we initialize a clothed human 3D model in the canonical space using a single image avatar creation method (Xiu et al. 2023), unlike SMPL+D initialization of other methods. Such a different starting point decreases the gap between initialization and subsequent deformation iterations, leading to higher accuracy of mapping canonical points to the current frame space (Chen et al. 2021; Zheng et al. 2022). During the iterations of reconstruction, we focus on the dynamics in the non-rigid field and carefully update weights in the skinning transformation field to better explain and simulate clothing movement. We propose a dynamics method DDF to model the influence of related forces on the deformation of loose clothing, which enhances deformation interpretability and enables DLCA-Recon to more accurately simulate real-world physics. Moreover, we optimize human poses and manage the overall network optimization to prevent training collapse. Our contributions could be summarized as follows,

- We propose to use an estimated human geometry as mesh initialization in the canonical space, which could better guide the SMPL weight propagation to the body and clothes. We especially fine-tune body pose and skinning weights to improve skinning transformation;
- We propose a dynamic deformation field (DDF) to account for all major contributive forces, which could effectively model the free movement of body and clothes;
- Extensively experimental evaluations on benchmark datasets and our captured monocular videos demonstrate that our method outperforms existing methods. We provide a more robust spatial-temporal reconstruction method for 3D dynamic avatars with loose clothing.

Related Work

Clothed Human Reconstruction from Single-View Image. Traditional human reconstruction often adopts a parametric model, $e.g.$ SMPL (Loper et al. 2015) or SCAPE (Anguelov et al. 2005) and only recover a naked 3D body (Joo, Simon, and Sheikh 2018; Kanazawa et al. 2018). Many methods use "SMPL+D" to represent 3D clothed humans (Alldieck et al. 2018, 2019a,b; Zhu et al. 2019; Ma et al. 2020; Xiang et al. 2020). However, this "body+offset" approach is not flexible enough to model loose clothing like dresses and skirts.

Recent methods introduce implicit representation to increase topological flexibility. PIFu and PIFuhd (Saito et al. 2019, 2020) extract pixel-aligned spatial features from images to implicit surface function. Two methods do not leverage knowledge of the human body structure, resulting in overfitting the body poses in training data. Consequently, they fail to generalize the 3D model to novel poses and produce shapes with broken or disembodied limbs.

To address these issues, several methods (Huang et al. 2020; He et al. 2021; Zheng et al. 2021a,b; Liao et al. 2023) combine parametric body models with implicit representations. To further generalize to unseen poses, ICON (Xiu et al. 2022) regresses shapes from locally queried features. These approaches enhance robustness to unseen poses but still have not enough generalization ability to various, especially loose, clothing topologies. Recently, ECON (Xiu et al. 2023) directly generates the clothed human from bilateral normal integration, enabling loose-fitting clothing reconstruction. But it tends to output bent legs and incorrect thickness of human.

As these methods only consider single-image reconstruction, they cannot produce temporally consistent results. The results can be wrong in other views. Moreover, these methods require a large amount of 3D scanned ground truth to ensure generalization capability.

Clothed Human Reconstruction from Monocular Video. Traditional methods require personalized rigged templates as prior and track the pre-defined human model based on 2D observations (Xu et al. 2018; Habermann et al. 2019, 2020). These methods require pre-scanning and manual rigging, unsuitable for lightweight applications. Some explicit methods (Alldieck et al. 2018; Guo et al. 2021; Casado-Elvira, Trinidad, and Casas 2022; Moon et al. 2022) omit personalized rigged templates but are still limited to a fixed resolution and topologies. Some methods (Pons-Moll et al. 2017; Tiwari et al. 2020; Xiang et al. 2022; Casado-Elvira, Trinidad, and Casas 2022) reconstruct the clothing as a separate layer over the body with high-quality 3D clothing supervision.

Some approaches introduce implicit methods to capture the details and facilitate 3D body reconstruction. NeuralBody (Peng et al. 2021b) represents dynamic human NeRF based on SMPL. HumanNeRF (Weng et al. 2022) extends articulated NeRF to improve novel view synthesis. NeuMan (Jiang et al. 2022b) further adds a scene NeRF model. These methods model the geometry with a density field, yielding low-fidelity and spatial-temporal inconsistent human reconstruction. SelfRecon (Jiang et al. 2022a) combines explicit and implicit representation to reconstruct temporally consistent 3D clothed humans, but it could not reconstruct humans in loose clothing and free motion. Vid2avatar (Guo et al. 2023) utilizes self-supervised scene decomposition to achieve temporally consistent human reconstruction, but it is not special for loose clothing.

Methodology

The proposed DLCA-Recon method is schematically illustrated in Figure 2. Like SelfRecon (Jiang et al. 2022a), we jointly optimize the explicit and implicit representation. Given a monocular video, we first randomly choose one frame and define the canonical human representation in both explicit mesh and implicit signed-distance field (SDF). During training, DLCA-Recon estimates the pose parameters for each frame and its two neighboring frames, which would be inputted into the forward deformation to deform the 3D human model from canonical space to each frame's space. It consists of the non-rigid dynamic deformation field and the optimized skinning transformation field. Our proposed non-rigid dynamic deformation field (DDF) aims to capture the movement of loose clothing and generate spatial-temporal coherent explicit meshes. Finally, we utilize mask loss to control the shape of explicit mesh and improve details through normal loss and color loss.

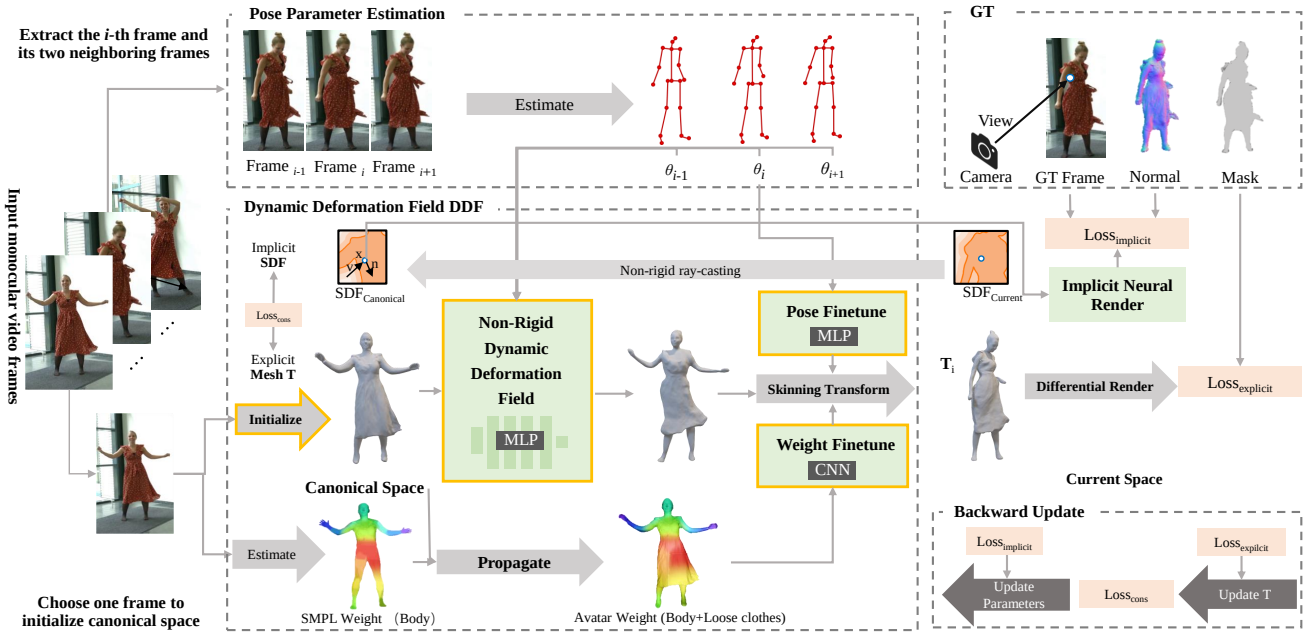

Figure 2: Diagram of DLCA-Recon. Inputting a monocular video, DLCA-Recon first initializes a 3D human mesh in canonical space. Through maintaining double representations of explicit mesh and implicit SDF for the avatar, DLCA-Recon handles the current *i*-th frame by pose parameter estimation, dynamic deformation field DDF, and loss calculation. In the backward update procedure, the explicit representation loss updates the mesh T in the canonical space, and then the consistency loss aligns the explicit representation and the implicit representation. Finally the implicit loss updates all learnable parameters. Gray arrow denotes an operation. Green rectangle represents a certain network and yellow border parts are our proposed modules.

Non-Parametric Initialization

Employing a coarse human body initialization provides enhanced adaptability to different types of loose clothing, so we use a coarse human representation instead of SMPL in canonical space. Specifically, we select a frame from the monocular video and feed it into ECON (Xiu et al. 2023) to obtain a geometry as the initialization. Then, to represent high-fidelity geometry, we define a canonical SDF S by an MLP f of the geometry using IGR (Gropp et al. 2020):

$$S = \{x_c \mid f(x_c) = 0\}, \tag{1}$$

where x_c is the vertex in canonical space.

Moreover, due to different cameras used in ECON and weak supervision on 2D images, the geometry estimated from ECON may have inconsistent scales and misaligned positions in canonical space. In 3D reconstruction, providing an initialization that is dimensionally and positionally incorrect could lead to convergence challenges, instability, and shape distortion. So it is necessary to predict an SMPL in canonical space, as well as align and scale the initialization with the SMPL.

Dynamic Deformation Field

The deformation of a clothed human cannot be fully represented by skinning transformation. Following prior works (Jiang et al. 2022a; Weng et al. 2022; Guo et al. 2023), we decompose the deformation field into a non-rigid deformation field and a skinning deformation field. Meanwhile, based on the force analysis of clothing, we propose a dynamic deformation field (DDF). Our new deformation field consists of a non-rigid dynamic deformation field and an initialization-based skinning transformation field. Given a monocular video depicting a clothed person in free motion, we generate per-frame SMPL (Loper et al. 2015) pose parameters $\{\theta_i | i = 1, ..., N\}$ using PyMAF (Zhang et al. 2021).

Non-Rigid Dynamic Deformation Field. According to dynamics, the motion of an object is related to the forces on it. Physics-based cloth simulation analyzes the internal and external forces acting on each vertex. We primarily address the free motion of the human in nature scenes. In this case, the forces acting on each vertex include gravity, traction, and friction resulting from human movement, as well as air resistance. In addition, internal force affects each other between the vertices. To simplify the representation, we compute the total force acting on a clothing vertex at time t_i using vector operations. According to Newton's second law, $F = ma$, the traction acting on a vertex at time t_i can be expressed as

$$\begin{aligned} F_{traction_i} &= ma_i = m\frac{\Delta v_i}{\Delta t} = m\frac{\Delta s_i}{\Delta t^2} \\ &= m\frac{x_{i+1} - x_{i-1}}{2\Delta t^2}, \end{aligned} \tag{2}$$

where a_i is the instantaneous acceleration and v_i is the instantaneous velocity at the current frame. The mass m and

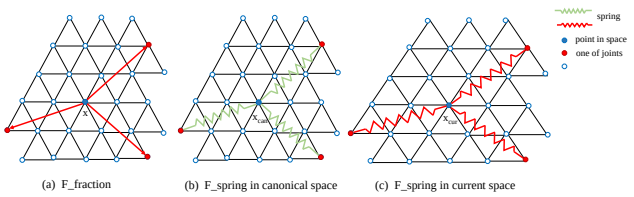

Figure 3: Forces on the clothing. We represent clothing as a combination of many triangular surfaces. Blue points on the triangles are points of the garment geometry, and red points indicate human joints. (a) shows the $F_{traction}$ that the point x receives from one of the joints. This force can be directly represented as a line from point x to the joint. (b) shows the state of the spring model in canonical space. We can simply understand that the green spring is in a relaxed state. In this case, the spring length sets the default to the original length. (c) shows the spring model in the current space. The red spring is stretched due to human movement.

time step Δt remain constant in a monocular video. Thus the force acting on a vertex at time t_i is determined by $x_{i+1}{-}x_{i-1}$. Since the traction force is mainly generated by human motion, $x_{i+1}{-}x_{i-1}$ can be expressed as a positive correlation of $J_{i+1}{-}J_{i-1}$. $J_{i+1}{-}J_{i-1}$ is the distance between the human joints of the subsequent video frames. Therefore, the traction can be presented as follows:

$$F_{traction_i} \propto (J_{i+1}{-}J_{i-1}). \tag{3}$$

Once the pose θ and translation T are available, the joints J can be known. So the traction is:

$$F_{traction_i} \propto (\theta_{i+1}, T_{i+1}, \theta_{i-1}, T_{i-1}). \tag{4}$$

The internal forces of the clothing primarily utilize a spring model, following Hooke's law $F_{spring} = -k\Delta x$. To calculate the internal force, we need to get Δx, which represents the spring's length change. As Figure 3 shows, in a macroscopic view, we use the vertex and human joints as the two ends of the spring. We assume the clothed human in canonical space represents an equilibrium state. So the initial length L of the spring can be viewed as the distance between vertices and force points, which can be considered as joints. Therefore, the displacement Δx_i of frame *i* is:

$$\begin{aligned} \Delta x_i &= x_i - L \\ &= (x_i - J_i) - (x_c - J_c) \\ &= (x_i - x_c) - (J_i - J_c). \end{aligned} \tag{5}$$

Since x_i is the part we need to solve, the formula can be presented as:

$$F_{spring_i} \propto (x_c, \theta_i, T_i, \theta_c, T_c), \tag{6}$$

where x_i is the vertex position at frame *i*, x_c is the vertex position in the canonical space.

Due to the uncontrollable scene, environmental factors such as wind may also affect the motion of the vertices. Therefore, a learnable variable φ is added as the effect of the environment on the vertices. As a result, we can get:

$$F_i \propto (x_c, \theta_{i+1}, T_{i+1}, \theta_{i-1}, T_{i-1}, \theta_i, T_i, \theta_c, T_c, \varphi). \tag{7}$$

We represent the non-rigid deformation of each frame with a learnable MLP. The point x' deformed by non-rigid deformation field can be represented as:

$$x' = F_i(x_c, \theta_{i+1}, T_{i+1}, \theta_{i-1}, T_{i-1}, \theta_i, T_i, \theta_c, T_c, \varphi). \tag{8}$$

Skinning Transformation Field. Given the *i*-th frame's pose parameter θ_i, we define a canonical-to-current space skinning transformation field. Following prior works (Jiang et al. 2022a; Lin et al. 2022b), we propagate the SMPL skinning weights in canonical space to get the initial skinning weights of arbitrary topologies. Specifically, we find the 30 nearest vertices on the SMPL mesh in canonical space for each point and average their skinning weights with IDW (inverse distance weight) as the initial weight.

Though using the initial skinning weights method to provide a good beginning, we still need to optimize them for the current subject. Similar to HumanNeRF (Weng et al. 2022), we employ a CNN to learn the weight offsets instead of solving for the entire set of skinning weights:

$$w = softmax(log(w_{init}) + CNN(x'; z)), \tag{9}$$

which can achieve faster and more accurate weight optimization.

We use PyMAF (Zhang et al. 2021) to extract SMPL parameters from images. PyMAF leverages a feature pyramid and rectifies the predicted parameters explicitly based on the mesh-image alignment status. Although PyMAF improves the alignment between meshes and images on 2D planes, it struggles to tackle the depth ambiguity problem in 3D space. To address this, we introduce an additional network to optimize the human pose. Similar to the skinning weight optimization, we use an MLP to obtain a relative value for pose optimization:

$$\Delta\Omega = MLP_\theta(\Omega), \tag{10}$$

where $\Omega = (\omega_0, ..., \omega_k)$ are local joint rotations represented as axis-angle vectors ω_i. We keep the joints J fixed and optimize the relative updates of the joint angles, $\Delta\Omega = (\Delta\omega_0, ..., \Delta\omega_k)$. We then apply these updates to Ω to obtain the updated rotation vectors:

$$pose(\theta) = (J, \Delta\Omega \otimes \Omega). \tag{11}$$

Finally, the skinning deformation is:

$$x_i = W(x', pose(\theta), w). \tag{12}$$

Delayed Optimization

During the optimization of the overall network, there are several modules and a lot of learnable parameters. As a result, the modules are not decoupled from each other. This leads to suboptimal learning outcomes for each module. Inspired by HumanNeRF (Weng et al. 2022), we deal with this issue by managing the overall optimization process. The optimizations for pose and skinning weights are disabled at the beginning of training, and they are gradually enabled when the non-rigid deformation network acquires a certain level of representation capacity. This approach can effectively alleviate the burden of network learning. Furthermore, pose optimization is only applied in the skinning transformation. Applying pose optimization in non-rigid transformation will increase the complexity of the non-rigid network and produce poor results.

Implicit Rendering Network

To obtain accurate geometry, we use surface rendering instead of volume rendering. While volume rendering can produce good rendering results, it generally yields poor geometry. Following the approach of IDR (Yariv et al. 2020), we input surface point, normal, view direction, and global geometric features into an MLP to estimate the colors of surface points. The obtained colors of points fully consider BRDF and global illumination and approximate the surface light field.

We only train the color field in canonical space to reduce memory usage and parameter amount. Followed by SelfRecon (Jiang et al. 2022a), we sample pixels within the ground truth mask and utilize non-rigid ray casting to obtain the corresponding point x_c in canonical space. In the meantime, we compute its normal $n_{x_c} = \bigtriangledown f(x_c)$ by gradient calculation. Given the camera information, we can determine the viewing direction v of each surface point x_d in the current space. By using the Jacobian matrix $J_{x_d}(x_c)$ of the deformation point $x_i = W(F_i(x_c))$, we can transform v to the viewing direction v_c of x_c in canonical space. Finally, we use an MLP to compute the color L_{x_c} of x_c, formulated as:

$$L_{x_c} = MLP_{color}(x_c, n_{x_c}, v_{x_c}). \tag{13}$$

Loss Function

During the computation of the explicit loss, we regard the canonical mesh T as an optimizable variable and compute its gradient together with the whole network. Then in the consistency loss, we connect explicit variations with the implicit representation. Explicit loss includes mask loss, while implicit losses include color loss, normal loss, and the Eikonal loss.

Mask Loss. We use the point cloud-based renderer in PyTorch3D and camera to render out the mask O_i' of the *i*-th frame mesh, and target mask O_i to calculate IoU loss:

$$loss_{IoU} = 1 - \frac{\|O_i' \otimes O_i\|_1}{\|O_i' \oplus O_i - O_i' \otimes O_i\|_1}, \tag{14}$$

where $\otimes$ and $\oplus$ are the operators that perform element-wise product and sum respectively.

Normal loss. We use the normal map predicted by PIFuHD (Saito et al. 2020) to refine the geometry. By gradient calculation, we can easily get normal n_{x_c}. In addition, we need to convert the corresponding predicted normal N from current space to canonical space, which can be calculated using $J_{x_d}(x_c)^T$. Therefore, there is a normal loss:

$$loss_{norm} = \left\| n_{x_c} - unit(J_{x_d}(x_c)^T N) \right\|_2, \tag{15}$$

where $unit(\cdot)$ means to normalize the vector.

Color loss. We minimize the color difference between the rendered image L_i and the input frame I_i as:

$$loss_{color} = |L_i - I_i|. \tag{16}$$

Eikonal Loss. We adopt the regular loss of IGR (Gropp et al. 2020) to make f to be signed distance function:

$$loss_{eik} = (\|\bigtriangledown f(x_c)\|_2 - 1)^2. \tag{17}$$

Subject	Normal MAE* ↓		Mask IoU ↑	
	SelfRecon	Ours	SelfRecon	Ours
Antonia	11.19	**6.14**	0.887	**0.904**
Magdalena	11.40	**7.69**	0.901	**0.912**
FranziRed	9.32	**5.29**	0.735	**0.912**
LCJ	11.38	**7.63**	0.884	**0.910**
LYZ	20.71	**8.00**	0.831	**0.919**
ZJ	15.74	**9.46**	0.819	**0.906**

Table 1: Quantitative comparison on geometry. "↑" indicates the higher the better, and "↓" indicates the lower the better. Normal MAE* = Normal MAE $\times 10^3$. "Antonia" and "Magdalena" are from DeepCap Dataset (Habermann et al. 2020), "FranziRed" is from DynaCap dataset (Habermann et al. 2021), and others are self-captured real sequences.

Finally, the implicit loss can be represented as:

$$Loss_{Implicit} = loss_{color} + loss_{eik} + \lambda loss_{norm}, \tag{18}$$

where $\lambda = 0.1$.

Consistency Loss. After explicit iteration, the canonical mesh T is updated to $\hat{T}$. To maintain consistent between the implicit SDF f and the updated explicit mesh $\hat{T}$ during implicit iteration, we employ a consistency loss from SelfRecon (Jiang et al. 2022a):

$$Loss_{cons} = \frac{1}{|\hat{T}|} \sum_{\hat{t} \in \hat{T}} \left| f(\hat{t}) \right| \tag{19}$$

where $\hat{t}$ is a vertex coordinate of $\hat{T}$. Intuitively, the loss demands alignment between $\hat{T}$ and the implicit surface.

Experiments

We evaluate our method on the DeepCap dataset (Habermann et al. 2020), DynaCap dataset (Habermann et al. 2021) and our own captured data (LCJ, LYZ and ZJ). Our data is captured in the wild with a static CANON EOS 6D MARK II camera. We fix the focal length and estimate the camera intrinsics using COLMAP. For each subject, we use 200-300 images for optimization. The optimization takes 200 epochs (about 48 hours) on a single NVIDIA RTX 3090 GPU.

Quantitative Evaluation

We utilize normal MAE and mask IoU as evaluation metrics of geometry. The results are presented in Table 1. We estimate the normal from PIFuHD (Saito et al. 2020) and the mask from RVM (Lin et al. 2022a) as ground truth. Table 1 shows that our reconstructed geometry outperforms SelfRecon in terms of both silhouettes and normal.

We report SSIM and PSNR to measure rendering quality. Results in Table 2 demonstrate that our method achieves higher accuracy than other methods under most metrics. Our method outperforms SelfRecon (Jiang et al. 2022a) and HumanNeRF (Weng et al. 2022) in all metrics. Compared to the SOTA method Vid2Avatar (Guo et al. 2023), our approach outperforms in loose clothing with a large wiggle amplitude.

Figure 4: Geometric qualitative comparison. From top to bottom: "Antonia" and "Magdalena" from DeepCap Dataset (Habermann et al. 2020), "FranziRed" from DynaCap dataset (Habermann et al. 2021), and others from self-captured real sequences.

Qualitative Evaluation

We also conduct qualitative comparisons with SelfRecon (Jiang et al. 2022a), SCARF (Feng et al. 2022) and Vid2Avatar (Guo et al. 2023) on the DeepCap Dataset (Habermann et al. 2020) and our own collected real sequences. More results are in the supplementary [1].

SelfRecon tends to produce physically incorrect body reconstructions. Due to the inherent limitations of NeRF, SCARF tend to reconstruct noisy geometry. SCARF proposes a hybrid model combining a mesh-based body with a NeRF-based clothing. Although it obtains relatively clean bodies, it reconstructs clothing with a lot of noise. This may be due to its inability to capture clothing dynamics in free motion. Following HumanNeRF and NeuMan (Jiang et al. 2022b), Vid2Avatar reconstructs an avatar via self-supervised scene decomposition. Though Vid2Avatar performs well on garments that are topologically similar to the body, it still fails to reconstruct loose clothing. It struggles to reconstruct loose-fitting clothing due to their fast dynamics. In contrast, our method generates complete and accurate results regardless of whether the clothing is tight or loose-fitting.

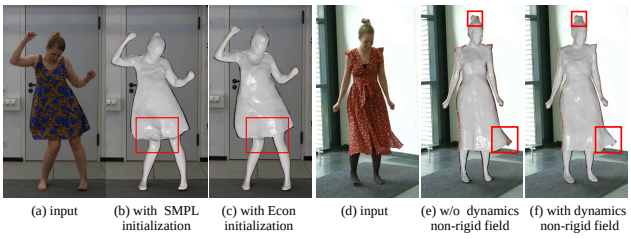

Figure 5: SMPL initialization cannot expand to clothes that are not similar to the body's topology (b). Econ initialization gives the correct topology for the clothed human (c). The dynamics non-rigid deformation improves clothing alignment and shape (d-f).

[1] https://github.com/surheaven/DLCA-Recon

Subject	PSNR ↑				SSIM ↑			
	SelfRecon	HumanNeRF	Vid2Avatar	Ours	SelfRecon	HumanNeRF	Vid2Avatar	Ours
Antonia	34.98	31.87	31.41	**37.98**	0.987	0.978	0.990	**0.991**
Magdalena	36.07	30.14	31.79	**39.77**	0.987	0.974	0.990	**0.992**
FranziRed	31.92	32.48	32.04	**34.09**	0.984	0.988	0.990	**0.991**
LCJ	29.17	31.62	34.13	**34.43**	0.989	0.986	**0.991**	**0.991**
LYZ	23.04	31.69	32.49	**38.14**	0.966	0.981	**0.987**	**0.987**
ZJ	31.50	32.00	35.70	**36.94**	0.979	0.984	**0.989**	0.986

Table 2: Quantitative comparison on rendering. "↑" indicates the higher the better, and "↓" indicates the lower the better. "Antonia" and "Magdalena" are from DeepCap Dataset (Habermann et al. 2020), "FranziRed" is from DynaCap dataset (Habermann et al. 2021), and others("LCJ", "LJZ", "ZJ") are self-captured real sequences.

	Normal MAE* ↓			IoU ↑			PSNR ↑			SSIM ↑		
	Antonia	Magdalena	Lab*	Antonia	Magdalena	Lab*	Antonia	Magdalena	Lab*	Antonia	Magdalena	Lab*
baseline	11.19	11.40	18.90	0.887	0.901	0.858	34.98	36.07	26.63	0.987	0.987	0.978
baseline+initialization	6.44	8.67	10.65	0.900	0.905	0.904	37.42	37.94	32.03	0.985	0.983	0.982
baseline+initialization+dynamic non-rigid	**6.03**	7.93	10.22	0.903	0.910	0.916	37.68	38.80	33.31	0.985	0.984	0.983
Ours (w/o dynamic non-rigid)	6.65	8.83	10.90	0.895	0.904	0.903	37.01	37.96	32.03	0.985	0.983	0.982
Ours (full model)	6.14	**7.69**	**10.22**	**0.904**	**0.912**	**0.921**	**37.98**	**39.77**	**33.81**	**0.990**	**0.992**	**0.985**

Table 3: Ablation study of geometry and rendering."↑" indicates the higher the better, and "↓" indicates the lower the better. Normal MAE* = Normal MAE $\times 10^3$. We compute averages over 3 sequences of Lab Dataset. Lab Dataset contains self-captured video clips. "Ours (w/o dynamic non-rigid)" means we use another non-rigid deformation field with only frame index. Our full model contains initialization, dynamic non-rigid and optimized skinning deformation fields, as well as a pose decoder.

Ablation Study

Effect of Initialization. To reconstruct human wearing various types of clothing, we select a frame from the video and employ ECON (Xiu et al. 2023) to derive an initial geometric. Table 3 shows the ablation experiments using ECON and SMPL initialization. As demonstrated in the results, when using SMPL initialization, even in conjunction with the SDF-based implicit method, it is prone to get stuck in local optima and cannot generate loose-fitting clothing. Figure 5 shows that the initialization of ECON lays the foundation for avatar reconstruction. Effective initialization gives the correct topology and facilitates network learning.

Effect of Dynamic Non-Rigid Deformation Field. Table 3 shows that the dynamic non-rigid field can better capture clothing movement, especially in the case of loose clothing. It proves the validity of force formulation in a non-rigid deformation field. Figure 5 visualizes the importance of including dynamics non-rigid motion. To verify the force formulation in the dynamic non-rigid field, we conduct an ablation study between dynamic non-rigid MLP and a non-rigid MLP with only frame index in Table 3.

Effect of Optimized Skinning Deformation Field. In Table 3, we find that using weights fine-tuning can get more correct geometry and improve the metrics to some extent. In a word, dynamics non-rigid alone is enough for significant improvement. Adding LBS fine-tuning provides further gains. We also conduct ablation experiments on the pose decoder and delayed optimization (see supplementary).

Effect of SMPL Pose Estimation Method. We replace PyMAF (Zhang et al. 2021) with a slightly non-robust method called SPIN (Kolotouros et al. 2019), which is compared in PyMAF work. Table 4 shows that our method can get stable accurate results with SPIN. As we employ a pose decoder to refine poses and apply a skeleton smoothness loss to maintain low-frequency joint motion trajectories, these measures decrease our sensitivity to SMPL poses.

	Normal MAE* ↓	IoU ↑	PSNR ↑	SSIM ↑
FranziRed(PyMAF)	5.29	0.912	34.09	0.991
FranziRed(SPIN)	5.31	0.912	34.04	0.991

Table 4: Ablation study of SMPL pose estimation method.

Conclusion and Discussion

In this paper, we have presented a method named DLCA-Recon to reconstruct 3D avatars from monocular in-the-wild videos. By employing force analysis in non-rigid deformations and optimizing skinning weights through initialization, we can effectively capture the free motion of bodies and clothes. Managing the overall network optimization process helps mitigate the coupling between modules to a certain extent. Without the requirement of scans as supervision, DLCA-Recon can reconstruct high-fidelity humans dressed in a variety of clothing styles from monocular videos in the wild.

DLCA-Recon still has several limitations. First, due to employing surface rendering loss, our method is limited by the accuracy of the ground truth mask. Second, current approaches rely on predicted normal maps to improve geometric details. Lastly, the geometry obtained from 2D supervision is still inferior to 3D supervision. We will address these issues in the future.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (No. 61972298 and No. 62372336), CAAI-Huawei MindSpore Open Fund and Wuhan University-Huawei GeoInformatices Innovation Lab.

References

Alldieck, T.; Magnor, M.; Bhatnagar, B. L.; Theobalt, C.; and Pons-Moll, G. 2019a. Learning to reconstruct people in clothing from a single RGB camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1175–1186.

Alldieck, T.; Magnor, M.; Xu, W.; Theobalt, C.; and Pons-Moll, G. 2018. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8387–8397.

Alldieck, T.; Pons-Moll, G.; Theobalt, C.; and Magnor, M. 2019b. Tex2shape: Detailed full human body geometry from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2293–2303.

Anguelov, D.; Srinivasan, P.; Koller, D.; Thrun, S.; Rodgers, J.; and Davis, J. 2005. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, 408–416.

Bao, Z.; Long, C.; Fu, G.; Liu, D.; Li, Y.; Wu, J.; and Xiao, C. 2022. Deep Image-based Illumination Harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18542–18551.

Cao, T.; Luo, F.; Fu, Y.; Zhang, W.; Zheng, S.; and Xiao, C. 2022. DGECN: A depth-guided edge convolutional network for end-to-end 6D pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3783–3792.

Cao, T.; Zhang, W.; Fu, Y.; Zheng, S.; Luo, F.; and Xiao, C. 2023. DGECN++: A Depth-Guided Edge Convolutional Network for End-to-end 6D Pose Estimation via Attention Mechanism. *IEEE Transactions on Circuits and Systems for Video Technology*.

Casado-Elvira, A.; Trinidad, M. C.; and Casas, D. 2022. PERGAMO: Personalized 3d garments from monocular video. In *Computer Graphics Forum*, volume 41, 293–304. Wiley Online Library.

Chen, X.; Zheng, Y.; Black, M. J.; Hilliges, O.; and Geiger, A. 2021. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11594–11604.

Feng, Y.; Yang, J.; Pollefeys, M.; Black, M. J.; and Bolkart, T. 2022. Capturing and animation of body and clothing from monocular video. In *SIGGRAPH Asia 2022 Conference Papers*, 1–9.

Gropp, A.; Yariv, L.; Haim, N.; Atzmon, M.; and Lipman, Y. 2020. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*.

Guo, C.; Chen, X.; Song, J.; and Hilliges, O. 2021. Human performance capture from monocular video in the wild. In *2021 International Conference on 3D Vision (3DV)*, 889–898. IEEE.

Guo, C.; Jiang, T.; Chen, X.; Song, J.; and Hilliges, O. 2023. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12858–12868.

Habermann, M.; Liu, L.; Xu, W.; Zollhoefer, M.; Pons-Moll, G.; and Theobalt, C. 2021. Real-time deep dynamic characters. *ACM Transactions on Graphics (ToG)*, 40(4): 1–16.

Habermann, M.; Xu, W.; Zollhoefer, M.; Pons-Moll, G.; and Theobalt, C. 2019. Livecap: Real-time human performance capture from monocular video. *ACM Transactions On Graphics (TOG)*, 38(2): 1–17.

Habermann, M.; Xu, W.; Zollhofer, M.; Pons-Moll, G.; and Theobalt, C. 2020. Deepcap: Monocular human performance capture using weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5052–5063.

He, T.; Collomosse, J.; Jin, H.; and Soatto, S. 2020. Geopifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. *Advances in Neural Information Processing Systems*, 33: 9276–9287.

He, T.; Xu, Y.; Saito, S.; Soatto, S.; and Tung, T. 2021. Arch++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11046–11056.

Huang, Z.; Xu, Y.; Lassner, C.; Li, H.; and Tung, T. 2020. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3093–3102.

Jiang, B.; Hong, Y.; Bao, H.; and Zhang, J. 2022a. Selfrecon: Self reconstruction your digital avatar from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5605–5615.

Jiang, W.; Yi, K. M.; Samei, G.; Tuzel, O.; and Ranjan, A. 2022b. Neuman: Neural human radiance field from a single video. In *European Conference on Computer Vision*, 402–418. Springer.

Joo, H.; Simon, T.; and Sheikh, Y. 2018. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8320–8329.

Kanazawa, A.; Black, M. J.; Jacobs, D. W.; and Malik, J. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7122–7131.

Kolotouros, N.; Pavlakos, G.; Black, M. J.; and Daniilidis, K. 2019. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2252–2261.

Li, Y.; Luo, F.; and Xiao, C. 2023. Monocular human depth estimation with 3D motion flow and surface normals. *The Visual Computer*, 39(8): 3701–3713.

Liao, T.; Zhang, X.; Xiu, Y.; Yi, H.; Liu, X.; Qi, G.-J.; Zhang, Y.; Wang, X.; Zhu, X.; and Lei, Z. 2023. High-Fidelity Clothed Avatar Reconstruction from a Single Image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8662–8672.

Lin, S.; Yang, L.; Saleemi, I.; and Sengupta, S. 2022a. Robust high-resolution video matting with temporal guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 238–247.

Lin, S.; Zhang, H.; Zheng, Z.; Shao, R.; and Liu, Y. 2022b. Learning implicit templates for point-based clothed human modeling. In *European Conference on Computer Vision*, 210–228. Springer.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6): 1–16.

Luo, F.; Zhu, Y.; Fu, Y.; Zhou, H.; Chen, Z.; and Xiao, C. 2023. Sparse RGB-D images create a real thing: A flexible voxel based 3D reconstruction pipeline for single object. *Visual Informatics*, 7(1): 66–76.

Ma, Q.; Yang, J.; Ranjan, A.; Pujades, S.; Pons-Moll, G.; Tang, S.; and Black, M. J. 2020. Learning to dress 3d people in generative clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6469–6478.

Moon, G.; Nam, H.; Shiratori, T.; and Lee, K. M. 2022. 3d clothed human reconstruction in the wild. In *European conference on computer vision*, 184–200. Springer.

Peng, S.; Dong, J.; Wang, Q.; Zhang, S.; Shuai, Q.; Zhou, X.; and Bao, H. 2021a. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14314–14323.

Peng, S.; Zhang, Y.; Xu, Y.; Wang, Q.; Shuai, Q.; Bao, H.; and Zhou, X. 2021b. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9054–9063.

Pons-Moll, G.; Pujades, S.; Hu, S.; and Black, M. J. 2017. ClothCap: Seamless 4D clothing capture and retargeting. *ACM Transactions on Graphics (ToG)*, 36(4): 1–15.

Saito, S.; Huang, Z.; Natsume, R.; Morishima, S.; Kanazawa, A.; and Li, H. 2019. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2304–2314.

Saito, S.; Simon, T.; Saragih, J.; and Joo, H. 2020. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 84–93.

Tiwari, G.; Bhatnagar, B. L.; Tung, T.; and Pons-Moll, G. 2020. Sizer: A dataset and model for parsing 3d clothing and learning size sensitive 3d clothing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, 1–18. Springer.

Weng, C.-Y.; Curless, B.; Srinivasan, P. P.; Barron, J. T.; and Kemelmacher-Shlizerman, I. 2022. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, 16210–16220.

Xiang, D.; Bagautdinov, T.; Stuyck, T.; Prada, F.; Romero, J.; Xu, W.; Saito, S.; Guo, J.; Smith, B.; Shiratori, T.; et al. 2022. Dressing avatars: Deep photorealistic appearance for physically simulated clothing. *ACM Transactions on Graphics (TOG)*, 41(6): 1–15.

Xiang, D.; Prada, F.; Wu, C.; and Hodgins, J. 2020. Monoclothcap: Towards temporally coherent clothing capture from monocular rgb video. In *2020 International Conference on 3D Vision (3DV)*, 322–332. IEEE.

Xiu, Y.; Yang, J.; Cao, X.; Tzionas, D.; and Black, M. J. 2023. ECON: Explicit Clothed humans Optimized via Normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 512–523.

Xiu, Y.; Yang, J.; Tzionas, D.; and Black, M. J. 2022. Icon: Implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13286–13296. IEEE.

Xu, W.; Chatterjee, A.; Zollhöfer, M.; Rhodin, H.; Mehta, D.; Seidel, H.-P.; and Theobalt, C. 2018. Monoperfcap: Human performance capture from monocular video. *ACM Transactions on Graphics (ToG)*, 37(2): 1–15.

Yariv, L.; Kasten, Y.; Moran, D.; Galun, M.; Atzmon, M.; Ronen, B.; and Lipman, Y. 2020. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33: 2492–2502.

Zhang, H.; Tian, Y.; Zhou, X.; Ouyang, W.; Liu, Y.; Wang, L.; and Sun, Z. 2021. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11446–11456.

Zheng, Y.; Abrevaya, V. F.; Bühler, M. C.; Chen, X.; Black, M. J.; and Hilliges, O. 2022. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13545–13555.

Zheng, Y.; Shao, R.; Zhang, Y.; Yu, T.; Zheng, Z.; Dai, Q.; and Liu, Y. 2021a. Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6239–6249.

Zheng, Z.; Yu, T.; Liu, Y.; and Dai, Q. 2021b. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 44(6): 3170–3184.

Zhu, H.; Zuo, X.; Wang, S.; Cao, X.; and Yang, R. 2019. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4491–4500.