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]Wuhan University, Wuhan, China

luochunjie@whu.edu.cn, luofei@whu.edu.cn, wangyusen@whu.edu.cn, zhaoenxu@whu.edu.cn, cxxiao@whu.edu.cn

Appendices

A. Implementation details

A.1 Corresponding Points between Canonical Space and Current Space.

We transform the point x_c from canonical space to current frame space (i-th frame) and get deformed points x_i. We need to find the corresponding points between two spaces to utilize color loss. Following SelfRecon (Jiang et al. 2022a), we use differentiable non-rigid ray-casting. Specifically, given a ray r with camera position c and ray direction v, we can get the first intersection p_i on the deformed mesh. Moreover, with the intersected triangle on the deformed mesh, we can find p_c's corresponding point p_c on the canonical mesh by consistent barycentric weights. With p_c as the good initialization of x_c, we deform point $x_i = D_i(x_c) = W(F_i(x_c))$ and x_i is the intersection point of the ray r and the current space SDF. The SDF in canonical space is f, and p_c is on the surface of canonical mesh, so we need to drive $f(p_c)$ to 0. Specifically, we solve p by:

$$\begin{aligned} x_c &= \min |f(p_c)| + \frac{\|(p_i - c) \times v\|_2}{\|p_i - c\|_2} \\ &= \arg\min_{p_c} |f(p_c)| + \frac{\|(D_i(p_c) - c) \times v\|_2}{\|(D_i(p_c) - c\|_2}. \end{aligned} \quad (1)$$

A.2 More Details of Initialization.

We can initialize from any frame of the video. Following NeuMan (Jiang et al. 2022b), we prefer the frame where the limbs are separated for initialization. It avoids collision when warping from canonical space to current space. In Figure 2, the arms and the body are merged in the initialization. It leads to incorrect initial weights. We can see that the weight of the hand (purple) is incorrectly spread to the skirt. Additionally, forward deformation cannot be effectively performed when limbs are not separated, resulting in geometric distortions.

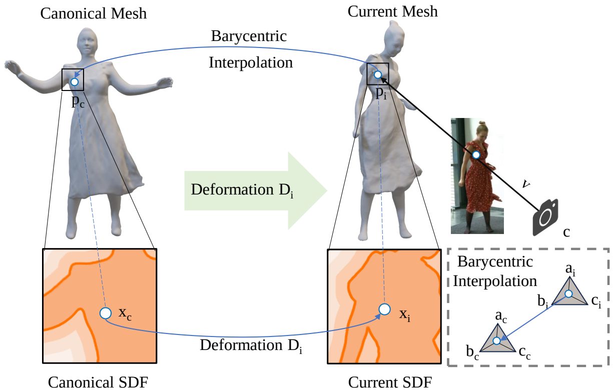


Figure 1: We render the mesh in the current space and find visible faces. We get the point p_i on one visible face and barycentric weights. Then we use the barycentric interpolation to initialize the corresponding point p_c in canonical space. $\triangle a_i b_i c_i$ is one visible face in the current space. Moreover, the $\triangle a_c b_c c_c$ is the corresponding face of $\triangle a_i b_i c_i$ in canonical space.

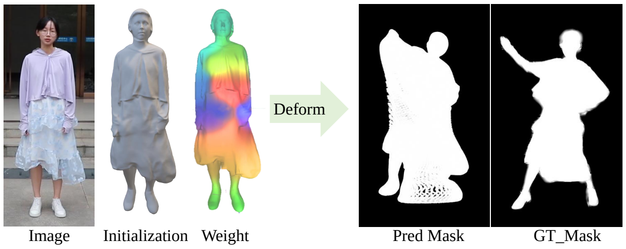


Figure 2: We select a frame for initialization. The initialization with limbs not separated leads to wrong initial weights and incorrect human body deformation.

A.3 Network Architecture.

As Figure 3 shows, the canonical human shape SDF network f includes 9 fully connected layers, each followed by a softplus activation layer. In Figure 4, the dynamic deformation field DDF consists of a dynamic non-rigid field and an optimized skinning transformation field. The non-rigid field comprises 5 linear layers, each of which is terminated by

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

a RELU activation. We apply LBS skinning transformation from SMPL (Loper et al. 2015) and optimize the skinning weights in the skinning transformation field. The weights optimization network starts with a fully-connected layer that transforms the latent code to a $1 \times 1 \times 1 \times 1024$ grid. Subsequently, it is combined with 6 transposed convolutions, progressively increasing the volume size while reducing channel count, and finally produces a volume of size $65 \times 225 \times 129 \times 24$. LeakyReLU is applied after MLP and transposed convolution layers. Pose finetune network input joint angles Ω into a 5-layer MLP with a width of 256 and output $\triangle\Omega$. The implicit rendering network is a 9-layer MLP finishing with a softplus activation layer.

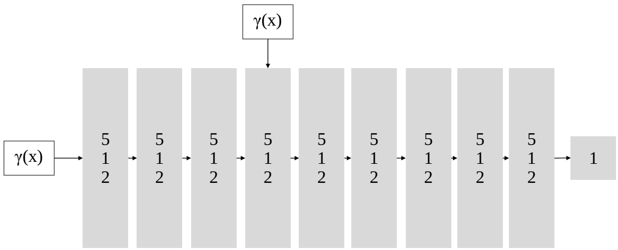


Figure 3: Canonical human shape SDF visualization.

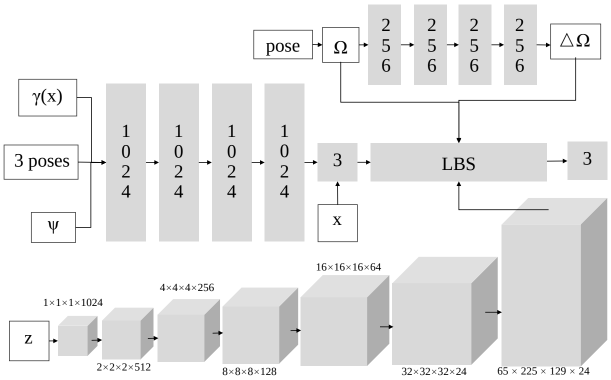


Figure 4: The visualization of dynamic deformation field DDF. DDF contains the dynamic non-rigid deformation field (middle), the pose decoder (top), and the weights finetune network (down).

B. More Results

B.1 Comparison with ECON.

In Figure 5, we conducted qualitative comparisons of Econ (Xiu et al. 2023). As we can see, ECON is limited by the normal and pose predictions, which makes it prone to obtaining incomplete and erroneous results. We also adopt an indirect experimental approach by comparing the mesh reconstruction when ECON and ours obtain similar and correct SMPL to ensure fairer comparisons. In contrast, our method can generate complete and correct clothed avatars.

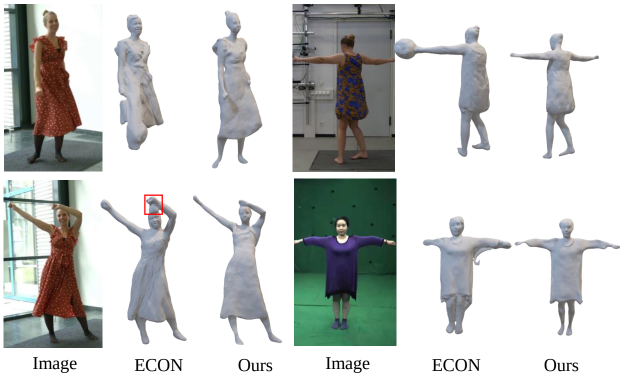


Figure 5: Geometric qualitative comparison of ECON. The first row demonstrates that ECON leads to reconstruction errors due to inaccurate SMPL pose estimation and mask splitting. The second row shows that, even with a similar and correct SMPL pose, ECON may still yield bad results.

B.2 Additional Qualitative Studies.

We also provide four views of estimated meshes in Figure 6. The unusual waist and arms result from pose errors caused by occlusion or invisibility. Existing pose estimation methods and 2D supervised avatar reconstruction methods cannot address this issue. Due to the 2D supervision inner limitation, it may cause some local unusual results.

We provide additional qualitative geometric results in Figure 7. We conducte experiments on types of clothing, and Figure 7 shows that DLCA-Recon can generate favorable results across a variety of clothing types.

B.3 Additional Quantitative Studies.

We make comparisons on public datasets. We compare with SCARF (Feng et al. 2022) and HumanNeRF (Weng et al. 2022) respectively. Table 1 and Table 2 show that we achieve better performance quantitatively.

Subject	PSNR ↑		SSIM ↑	
	SCARF	Ours	SCARF	Ours
male-3-casual	30.59	**37.30**	0.977	**0.984**
male-4-casual	31.79	**38.90**	0.970	**0.980**
female-3-casual	30.14	**42.63**	0.977	**0.987**
female-4-casual	29.96	**38.53**	0.972	**0.976**

Table 1: Quantitative comparison of SCARF in PeopleSnapshot dataset (Alldieck et al. 2018). We directly copy the metrics from SCARF. "↑" indicates the higher the better.

We present the comparison result with SOTA methods on both parameters' count and inference time in Table 3. When we get the best reconstruction results, we use a CNN to optimize the deformation field, which increases the parameters' count. However, it does not increase the inferring time a lot.

B.4 Additional Ablation Studies.

Effect of Pose Decoder. As shown in Figure 8, pose decoder can alleviate the impact of errors in the estimated pose

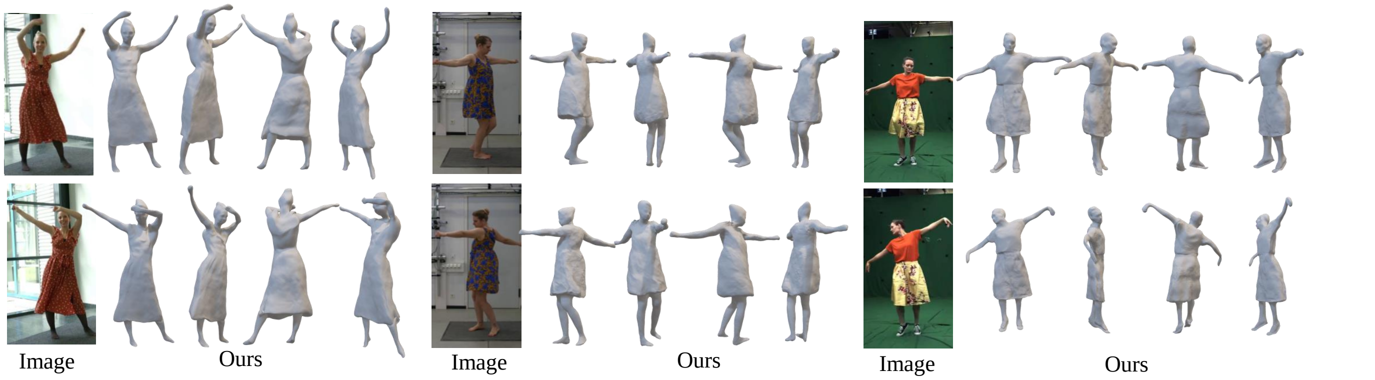


Figure 6: We show four perspectives of the clothed human. From left to right: “Antonia” and “Magdalena” from DeepCap Dataset (Habermann et al. 2020), “FranziRed” from DynaCap dataset (Habermann et al. 2021).

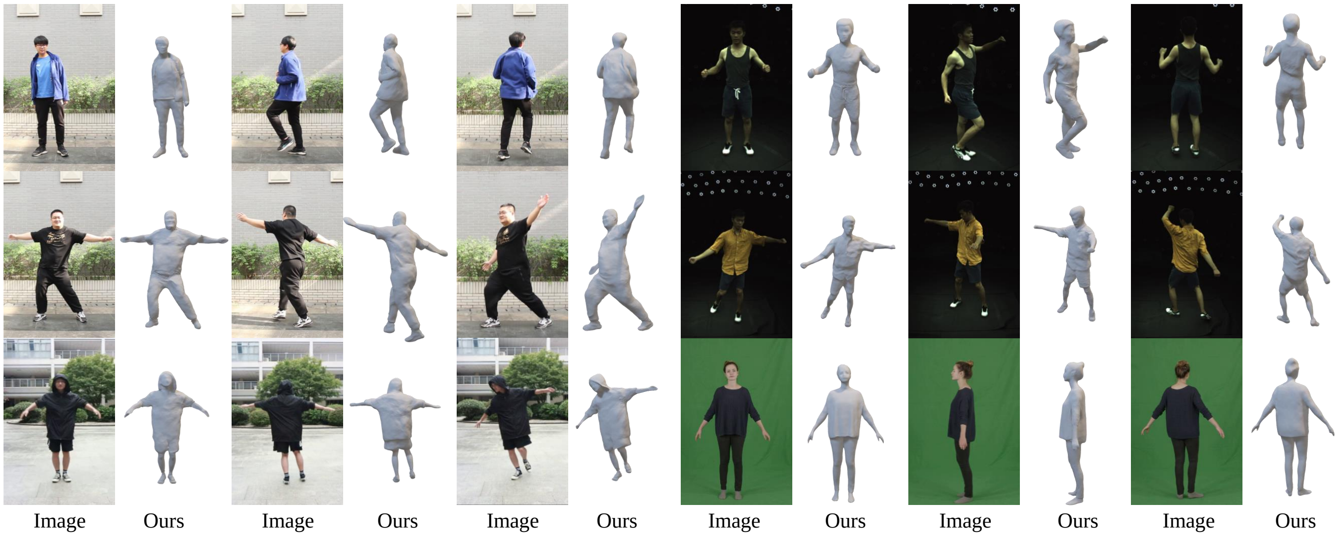


Figure 7: Additional qualitative geometric results. We have experiments on various types of clothing and DLCA-Recon consistently generates high-quality results. Each group shows images of the video and corresponding reconstructions. The results on the left are self-captured videos, the top two rows on the right are subject 377 and 393 from ZJU-MoCap dataset (Peng et al. 2021), and the bottom row on the right are “female-3-casual” from PeopleSnapshot dataset (Alldieck et al. 2018).

Subject	PSNR ↑		SSIM ↑	
	HumanNeRF	Ours	HumanNeRF	Ours
ZJU-MoCap dataset*	30.24	**31.10**	0.974	**0.978**

Table 2: Quantitative comparison of HumanNeRF on ZJU-MoCap dataset (Peng et al. 2021). We directly copy the metrics from HumanNeRF. We only compute “377”, “386”, “387”, “392”, “393” and “394” from ZJU-MoCap dataset. “↑” indicates the higher the better.

from PyMAF (Zhang et al. 2021). By using accurate poses, we can achieve more precise reconstruction results.

Effect of Delayed Optimization. Figure 9 demonstrates that without delayed optimization, the reconstruction method fails due to the excessive number of learnable parameters.

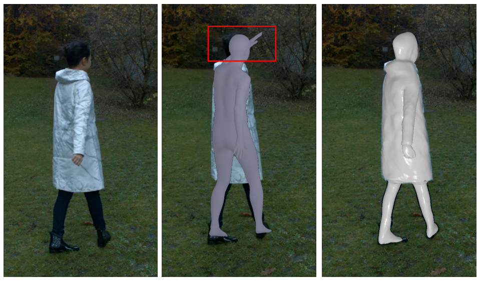
Methods	Params(M)	Infer Time(s)
SCARF(NeRF-based Method)	21.23	16.88
Vid2Avatar	1.42	992.98
SelfRecon	3.79	66.51
Ours	73.45	80.35

Table 3: Computation complexity comparison on methods.

References

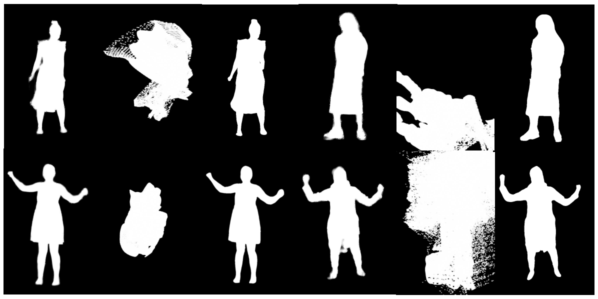
Alldieck, T.; Magnor, M.; Xu, W.; Theobalt, C.; and Pons-Moll, G. 2018. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8387–8397.

Feng, Y.; Yang, J.; Pollefeys, M.; Black, M. J.; and Bolkart, T. 2022. Capturing and animation of body and clothing from



(a) input (b) w/o pose decoder (c) with pose decoder

Figure 8: Pose decoder refines the body pose during optimization. It corrects the left arm from (b) to (c).



(a) input (b) w/o delayed optimization (c) with delayed optimization (a) input (b) w/o delayed optimization (c) with delayed optimization

Figure 9: Without delayed optimization, Geometries in (b) become distorted during the training process.

monocular video. In *SIGGRAPH Asia 2022 Conference Papers*, 1–9.

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.; 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.

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.

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.

Peng, S.; Zhang, Y.; Xu, Y.; Wang, Q.; Shuai, Q.; Bao, H.; and Zhou, X. 2021. 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.

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.

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.

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.