3D Human Reconstruction

PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization

Shunsuke Saito, Tomas Simon, Jason Saragih, Hanbyul Joo
CVPR 2020

Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans

Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, Xiaowei Zhou
CVPR 2021 Best paper candidate.

Lizi Fu
03/08/2022

PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization

- Motivation
- Goal
- Method
- Experiments
- Conclusion

PIFuHD — Motivation

- Fail to produce reconstructions with the level of detail often presented in the input images.

- Previous approaches tend to take low resolution images as input to cover large spatial context, and produce less precise (or low resolution) 3D estimates as a result

PIFuHD — Goal

Figure 1.1

- Goal: Achieve high-fidelity 3d reconstruction of clothed humans from a single image at a resolution sufficient to recover detailed information such as fingers, facial features and clothing folds.
- Input: single image
- Output: 3d reconstruction of clothed humans, such as fingers, facial features and clothing folds.

PIFuHD — Method

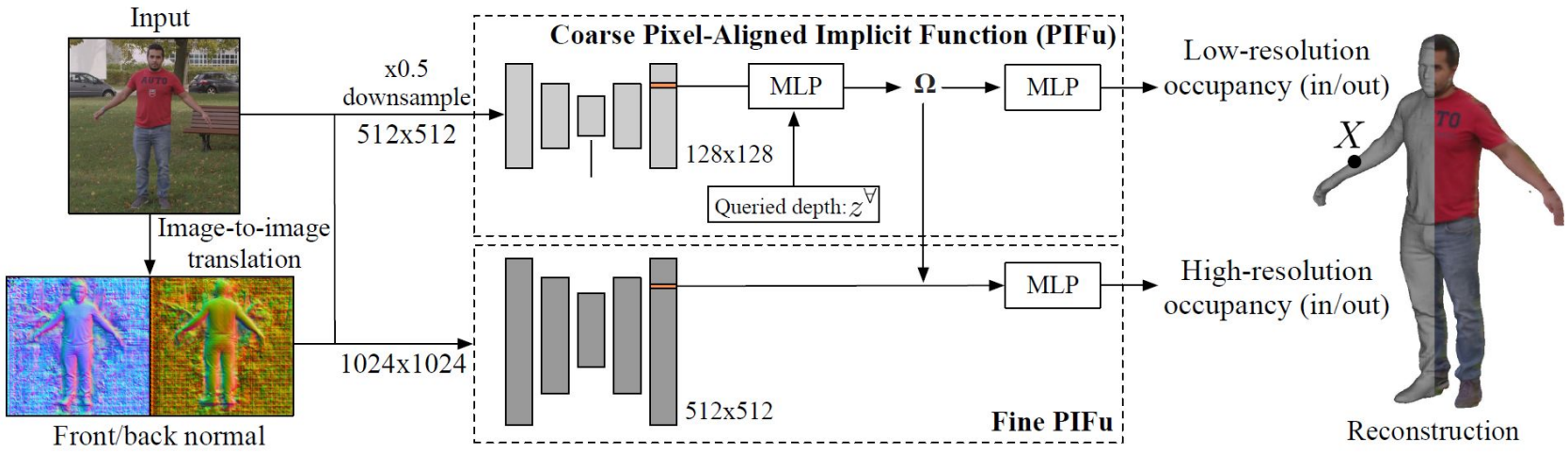

Figure 1.2 Overview of the framework. Two levels of pixel-aligned predictors produce high-resolution 3D reconstructions. The coarse level (top) captures global 3D structure, while high-resolution detail is added by the fine level.

PIFuHD — Method

- Pixel-Aligned implicit function

$$f(\mathbf{X}, \mathbf{I}) = \begin{cases} 1 & \text{if } \mathbf{X} \text{ is inside mesh surface} \\ 0 & \text{otherwise,} \end{cases} \qquad (1)$$

- Extracted function to estimates the occupancy of the query 3D point *X*

$$f(\mathbf{X}, \mathbf{I}) = g\left(\Phi\left(\mathbf{x}, \mathbf{I}\right), Z\right), \qquad (2)$$

PIFuHD — Method

Multi-level Pixel-Aligned implicit function

- Low-resolution occupancy

$$f^L(\mathbf{X}) = g^L\left(\Phi^L\left(\mathbf{x}_L, \mathbf{I}_L, \mathbf{F}_L, \mathbf{B}_L, \right), Z\right), \qquad (3)$$

- High-resolution occupancy

$$f^H(\mathbf{X}) = g^H\left(\Phi^H\left(\mathbf{x}_H, \mathbf{I}_H, \mathbf{F}_H, \mathbf{B}_H, \right), \Omega(\mathbf{X})\right), \qquad (4)$$

PIFuHD — Method

Binary Cross Entropy (BCE) loss function

$$\mathcal{L}_o = \sum_{\mathbf{X} \in \mathcal{S}} \lambda f^*(\mathbf{X}) \log f^{\{L,H\}}(\mathbf{X}) + (1-\lambda)\left(1 - f^*(\mathbf{X})\right) \log\left(1 - f^{\{L,H\}}(\mathbf{X})\right), \quad (5)$$

S denotes the set of samples at which the loss is evaluated,
λ is the ratio of points outside surface in ***S***,
$f^*(.)$denotes the ground truth occupancy at that location,
$f\{L,H\}g(.)$ are each of the pixel-aligned implicit functions of Low-resolution and High-resolution.

PIFuHD — Experiments

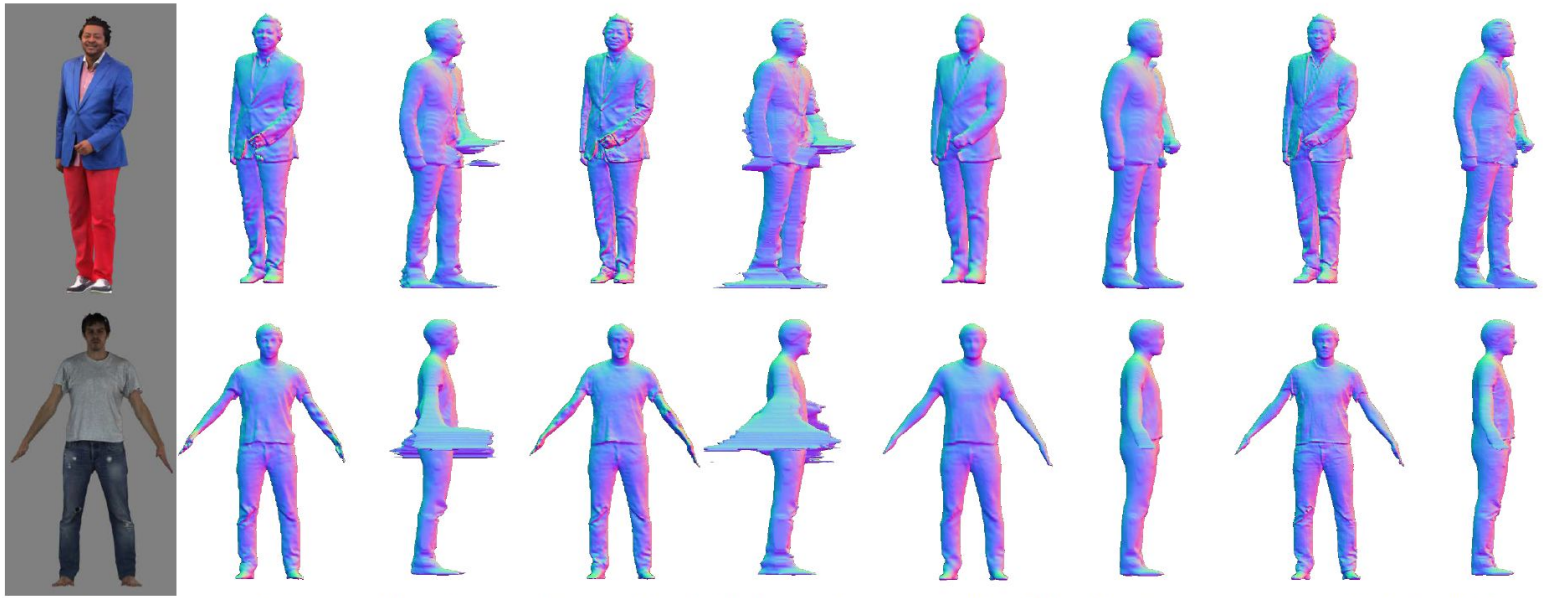

Figure 1.3 Qualitative evaluation of multi-level pixel-aligned implicit function on samples from RenderPeople and BUFF datasets. We compare the results of the final method with the results of other alternative designs.

PIFuHD — Experiments

Methods	RenderPeople			Buff		
	Normal	P2S	Chamfer	Normal	P2S	Chamfer
Fine module only	0.213	4.15	2.77	0.229	3.63	2.67
Fine module + Global image feature	0.165	2.92	2.13	0.183	2.767	2.24
Single PIFu	**0.109**	1.45	1.47	0.134	1.68	1.76
Ours (ML-PIFu, end-to-end)	0.117	1.66	1.55	0.147	1.88	1.81
Ours (ML-PIFu, alternate)	0.111	**1.41**	**1.44**	**0.133**	**1.63**	**1.73**
Ours with normals	0.107	1.37	1.43	0.134	1.63	1.75

Table 1.1 Quantitative evaluation on RenderPeople and BUFF datasets for single-view reconstruction. Units for point-tosurface and Chamfer distance are in cm.

PIFuHD — Experiments

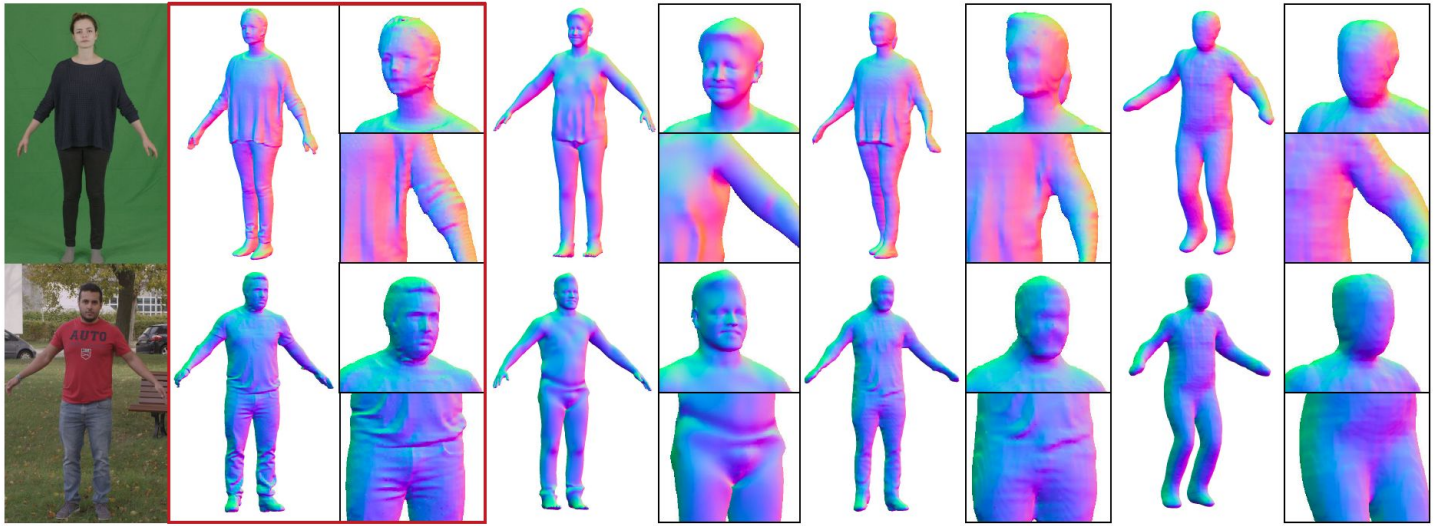

Figure 1.4 Qualitatively compare the method with state-of-the-art methods, including (c) Tex2shape, (d) PIFu, and (e) DeepHuman, on the People Snapshot dataset. By fully leveraging high-resolution image inputs, (b) our method can reconstruct higher resolution geometry compared to the existing methods.

PIFuHD — Experiments

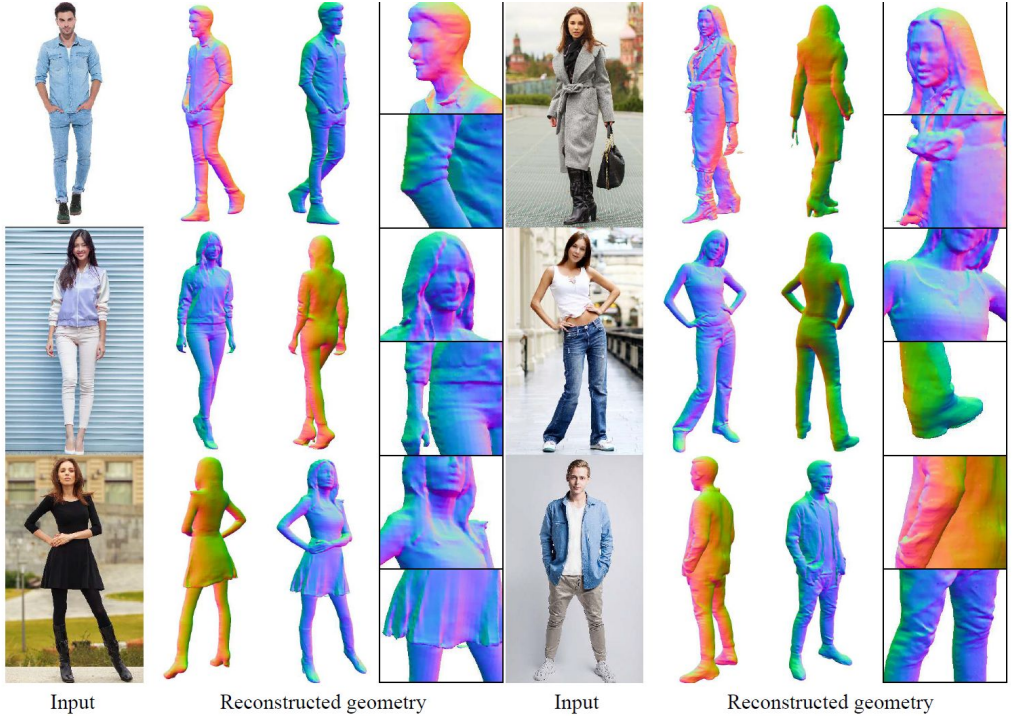

Figure 1.4 Qualitative results on Internet photos.

PIFuHD — Conclusion

- Present a multi-level framework t to arrive at high-resolution 3D reconstructions of clothed humans from a single image
- propagating global context through a scale pyramid as an implicit 3D embedding

Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans

- Motivation
- Goal
- Method
- Experiments
- Conclusion

Neural Body — Motivation

- A new approach capable of synthesizing photorealistic novel views of a performer in **complex motions** from a **sparse multi-view video**
- A novel implicit neural representation for a **dynamic human**, which enables us to effectively incorporate observations over video frames.
- The learned neural representations at different frames **share the same set of latent codes** anchored to a deformable mesh.

Neural Body — Goal

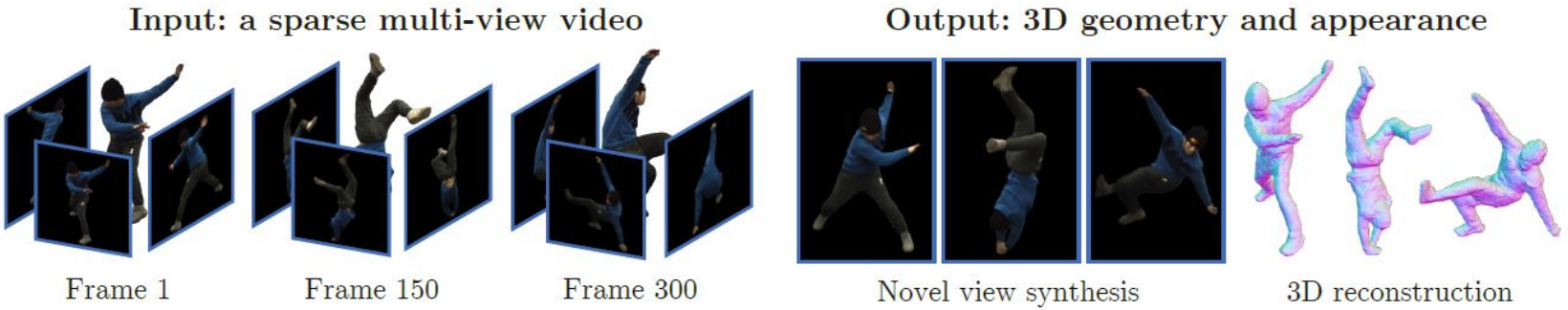

Figure 1

- Goal: View synthesis and body reconstruction for a human performer from a very sparse set of camera views.
- Input: Sparse multi-view videos
- Output: 3D geometry and apperance Novel view synthesis and 3D reconstruction

Neural Body — Method

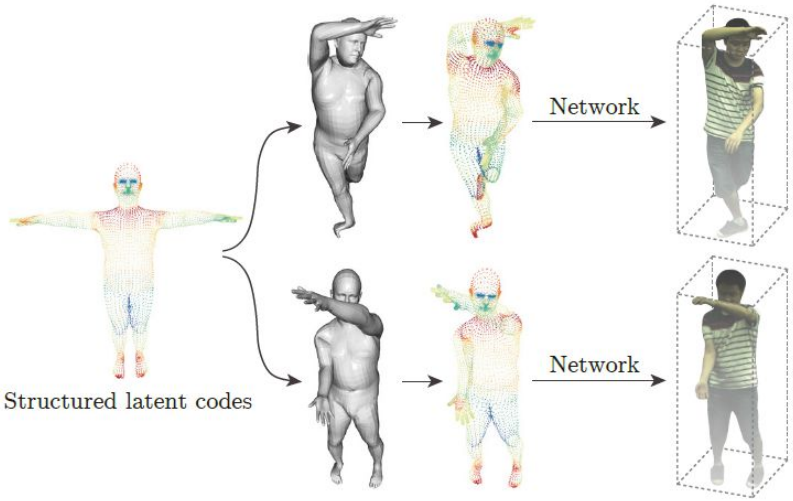

Figure 2.2 The basic idea of Neural Body

Neural Body — Method

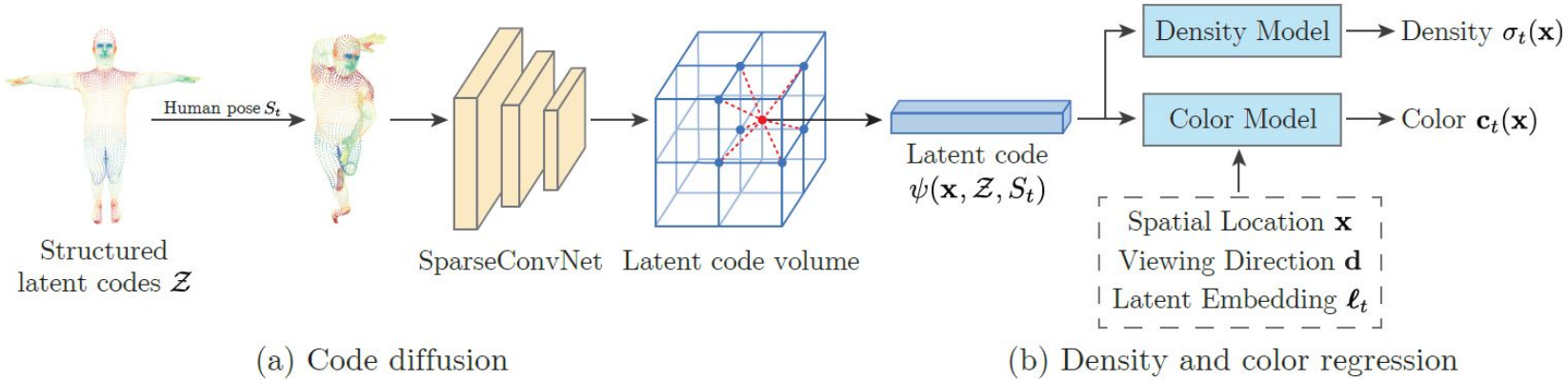

Figure 2.3 Implicit neural representation with structured latent codes

Neural Body — Method

- Density model

$$\sigma_t(\mathbf{x}) = M_\sigma(\psi(\mathbf{x}, \mathcal{Z}, S_t)), \qquad (1)$$

- Color model

$$\mathbf{c}_t(\mathbf{x}) = M_\mathbf{c}(\psi(\mathbf{x}, \mathcal{Z}, S_t), \gamma_\mathbf{d}(\mathbf{d}), \gamma_\mathbf{x}(\mathbf{x}), \boldsymbol{\ell}_t), \qquad (2)$$

Neural Body — Method

Volume Rendering

$$\tilde{C}_t(\mathbf{r}) = \sum_{k=1}^{N_k} T_k(1 - \exp(-\sigma_t(\mathbf{x}_k)\delta_k))\mathbf{c}_t(\mathbf{x}_k), \quad (3)$$

$$\text{where} \quad T_k = \exp(-\sum_{j=1}^{k-1} \sigma_t(\mathbf{x}_j)\delta_j), \quad (4)$$

Neural Body — Method

Optimize

Minimize the rendering error of observed images

$$\underset{\{\boldsymbol{\ell}_t\}_{t=1}^{N_t},\mathcal{Z},\Theta}{\text{minimize}} \sum_{t=1}^{N_t}\sum_{c=1}^{N_c} L(\mathcal{I}_t^c, P^c; \boldsymbol{\ell}_t, \mathcal{Z}, \Theta), \qquad (5)$$

L is the Loss function:

$$L = \sum_{\mathbf{r}\in\mathcal{R}} \left\| \tilde{C}(\mathbf{r}) - C(\mathbf{r}) \right\|_2^2, \qquad (6)$$

Neural Body — Experiments

Dataset

- ZJU-MoCap

  a multi-view dataset

- People-Snapshot dataset

  Monocular video dataset

Neural Body — Experiments

Novel view synthesis

	PSNR				SSIM			
	NV [37]	NT [61]	NHR [64]	OURS	NV [37]	NT [61]	NHR [64]	OURS
Twirl	22.09	25.78	26.68	**30.56**	0.831	0.929	0.935	**0.971**
Taichi	18.57	19.44	19.81	**27.24**	0.824	0.869	0.874	**0.962**
Swing1	22.88	24.96	24.73	**29.44**	0.726	0.905	0.902	**0.946**
Swing2	22.08	24.84	25.01	**28.44**	0.843	0.903	0.906	**0.940**
Swing3	21.29	23.50	23.47	**27.58**	0.842	0.896	0.894	**0.939**
Warmup	21.15	23.74	23.79	**27.64**	0.842	0.917	0.918	**0.951**
Punch1	23.21	24.93	25.02	**28.60**	0.820	0.877	0.879	**0.931**
Punch2	20.74	22.44	22.88	**25.79**	0.838	0.888	0.891	**0.928**
Kick	22.49	24.33	23.72	**27.59**	0.825	0.881	0.873	**0.926**
average	21.39	23.77	23.90	**28.10**	0.821	0.896	0.897	**0.944**

Table 2.1 Results on the ZJU-MoCap dataset in terms of PSNR and SSIM (higher is better). “NV” means Neural Volumes, and “NT” means Neural Textures.

Neural Body — Experiments

Novel view synthesis

Figure 2.4 Novel view synthesis on the ZJU-MoCap dataset.

Neural Body — Experiments

Novel view synthesis

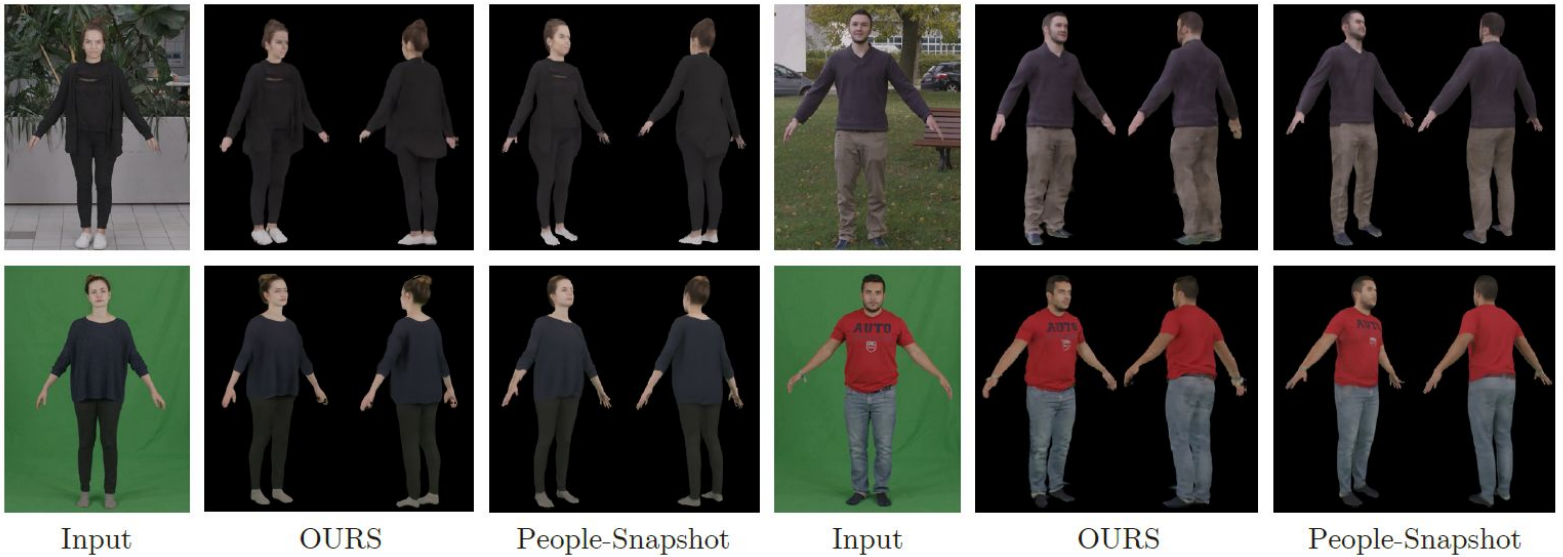

Figure 2.5 Novel view synthesis on monocular videos (People-Snapshot Dataset)

Neural Body — Experiments

3D Reconstruction

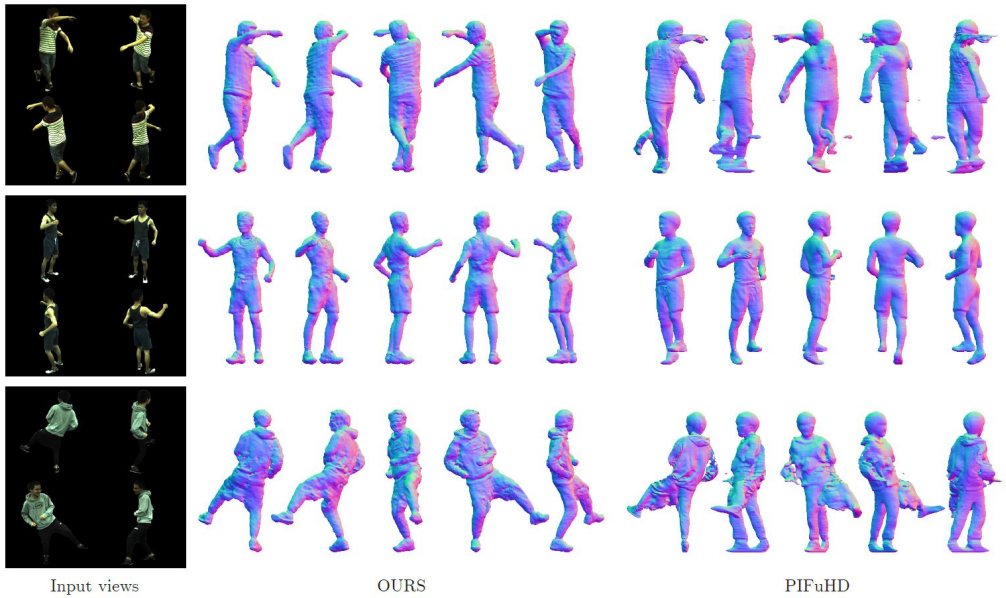

Figure 2.6 3D reconstruction on the ZJU-MoCap dataset

Neural Body — Experiments

3D Reconstruction

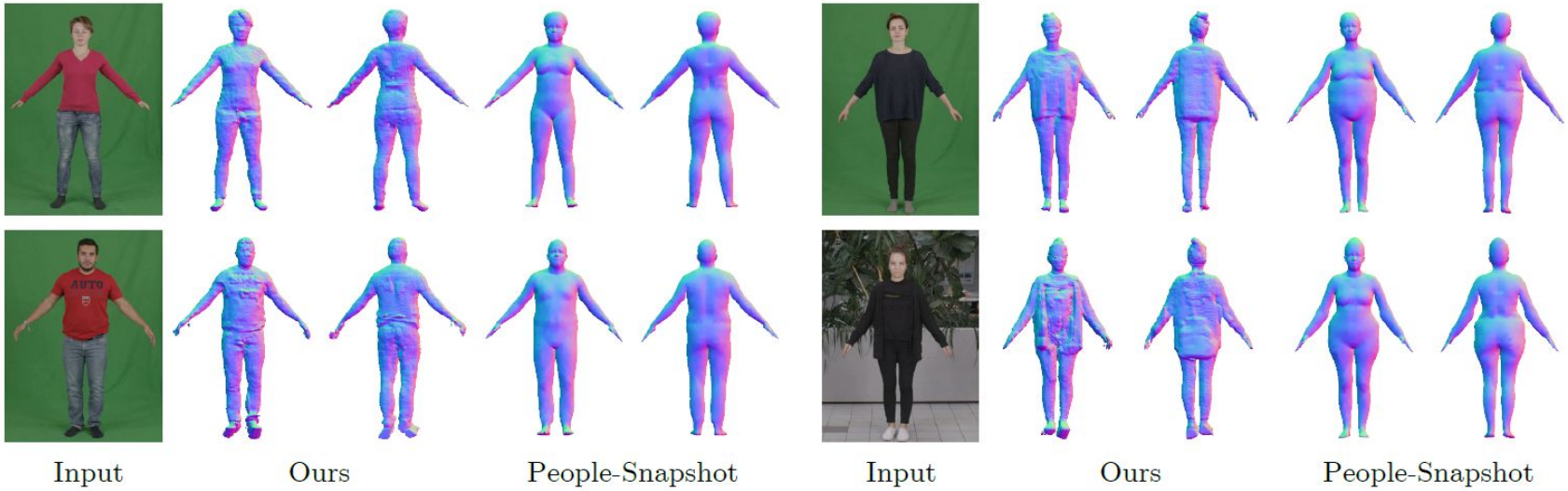

Figure 2.7 3D reconstruction on monocular videos(People-Snapshot).

Neural Body — Conclusion

- Introduce a novel implicit neural representation, named Neural Body, for novel view synthesis of dynamic humans from sparse multi-view videos.
- Establish a latent variable model that generates implicit fields at different video frames from the same set of latent codes
- Create a multi-view dataset called ZJU-MoCap that captures dynamic humans in complex motions and perform well.

Thanks for Watching

Lizi Fu
03/08/2022