ARCH: Animatable Reconstruction of Clothed Humans

Zeng Huang[1,2], Yuanlu Xu[1], Christoph Lassner[1], Hao Li[2], Tony Tung[1] [1]Facebook Reality Labs, Sausalito [2]University of Southern California

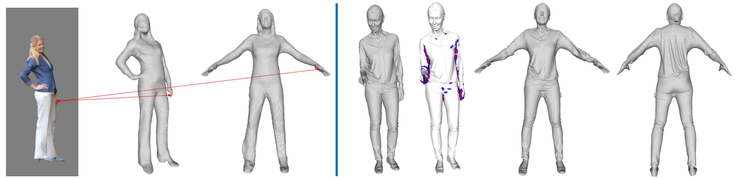
CVPR SEATTLE WASHINGTON JUNE 16-18 2020

Overview

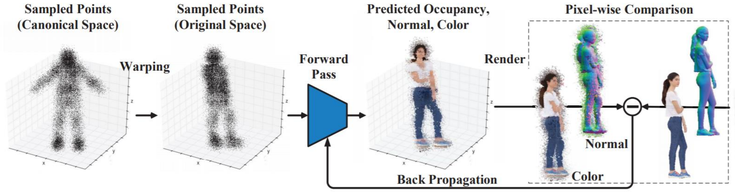
ARCH is an end-to-end framework for accurate reconstruction of **animation-ready** 3D clothed humans from a **monocular** image.

- ARCH is a learned pose-aware model that produces detailed 3D rigged full-body human avatars from a single unconstrained RGB image.
- A Semantic Space and a Semantic Deformation Field are created using a parametric 3D body estimator. They allow the transformation of 2D/3D clothed humans into a canonical space, reducing ambiguities in geometry caused by pose variations and occlusions in training data.
- Detailed surface geometry and appearance are learned using an implicit function representation with spatial local features.
- Furthermore, we propose additional per-pixel supervision on the 3D reconstruction using opacity-aware differentiable rendering.

Our Method uses a fitted template model, trains and predicts in **canonical space,** yielding **animation-ready** model:

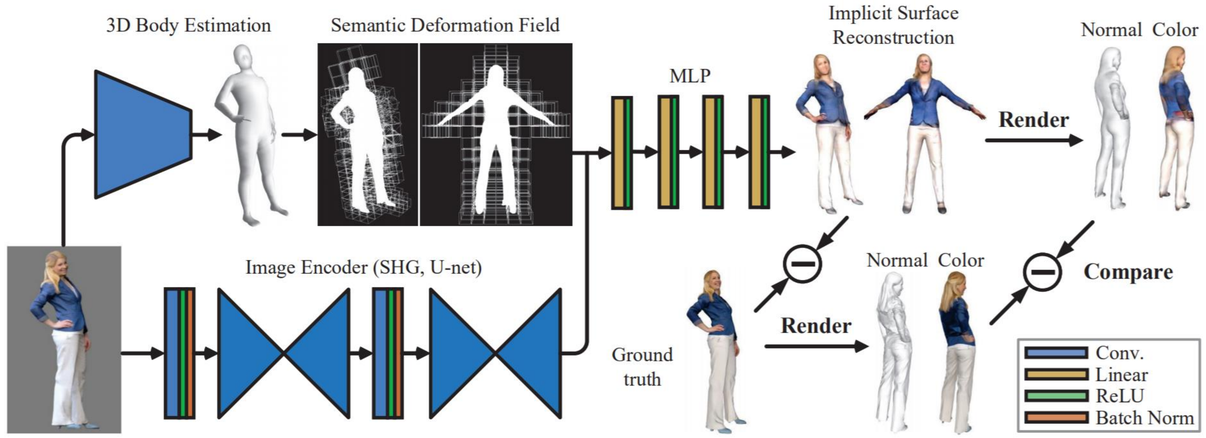


We utilize a **differentiable renderer** for additional supervision of predictions:



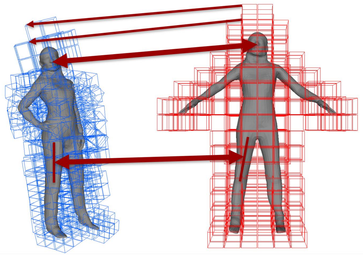
ARCH increases the fidelity of the reconstructed humans. We obtain more than 50% lower reconstruction errors for standard metrics compared to state-of-the-art methods on public datasets.

Input | Geometry Reconstruction | Normal Reconstruction | Color Reconstruction



Semantic Deformation Field

- Dense mapping from canonical space to object space with per-point body model skinning weight (a.k.a. Semantic Space)
- Pixel-aligned implicit representation
- Rigged reconstruction

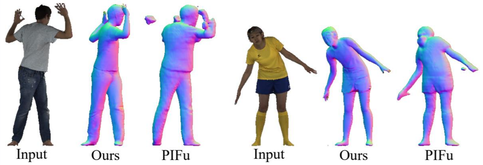


Experiments

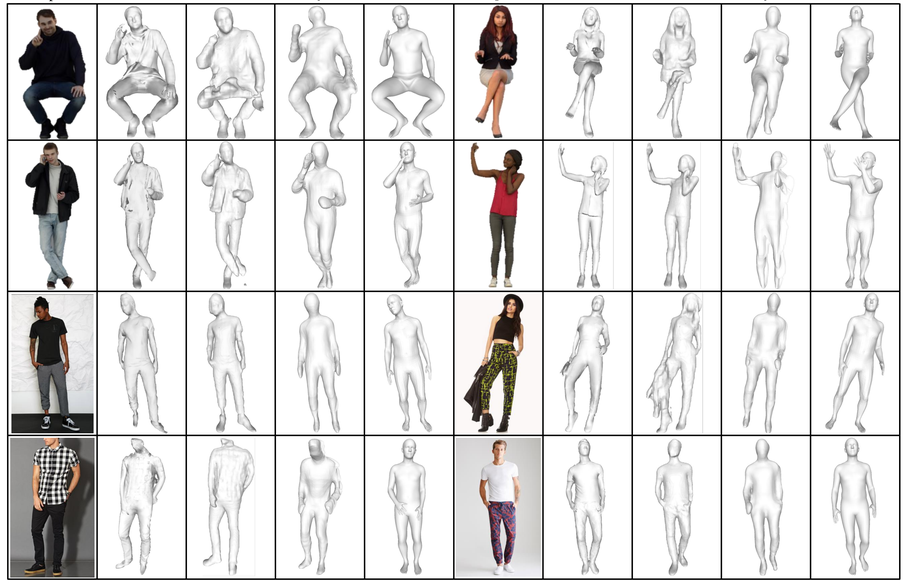
ARCH (Ours) robustly handles arbitrary poses with self-contacts and occlusions, and reconstructs higher level of details than existing methods. Outputs are rigged and animation-ready.
3D body estimation is performed by DenseRaC. We evaluate on 3D scan datasets and Internet images.

Methods	RenderPeople			BUFF		
	Normal	P2S	Chamfer	Normal	P2S	Chamfer
BodyNet [44]	0.26	5.72	5.64	0.31	4.94	4.52
SiCloPe [32]	0.22	3.81	4.02	0.22	4.06	3.99
IM-GAN [10]	0.26	2.87	3.14	0.34	5.11	5.32
VRN [17]	0.12	1.42	1.6	0.13	2.33	2.48
PIFu [40]	0.08	1.52	1.50	0.09	1.15	1.14
ARCH, baseline	0.080	1.98	1.85	0.081	1.74	1.75
+ SemDF	0.042	**0.74**	**0.85**	0.045	**0.82**	**0.87**
+ GRaC	**0.038**	**0.74**	**0.85**	**0.040**	**0.82**	**0.87**

Input | Animated Avatar



Input | Ours | PIFu | Input | Ours | PIFu



Reference:
[1] PIFu – Saito et al, "PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization", ICCV 19.
[2] BodyNet –Varol et al, "Bodynet: Volumetric inference of 3d human body shapes", ECCV18.
[3] HMR – Kanazawa et al, "End-to-end recovery of human shape and pose", CVPR18.
[4] DenseRaC – Xu et al, "Joint 3D Pose and Shape Estimation by Dense Render-and-Compare", ICCV19