MetaAvatar: Learning Animatable Clothed Human Models from Few Depth Images

Shaofei Wang[1], Marko Mihajlovic[1], Qianli Ma[1,2], Andreas Geiger[2,3], Siyu Tang[1]
[1]ETH Zürich, [2]Max Planck Institute for Intelligent Systems, Tübingen, [3]University of Tübingen

https://neuralbodies.github.io/metavatar/

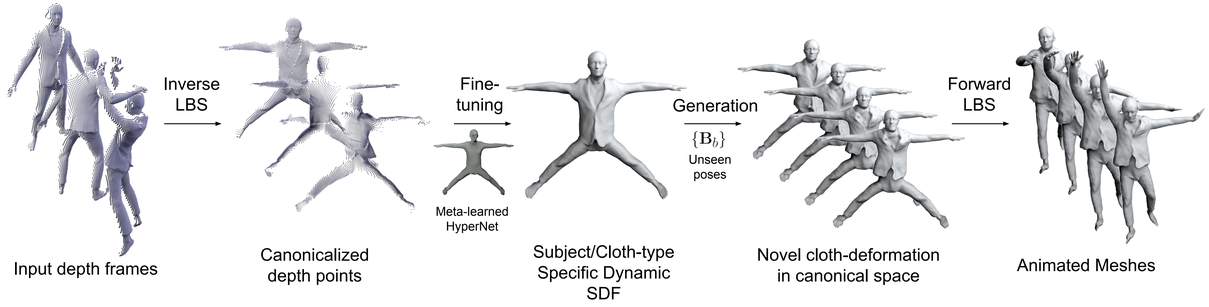
NEURAL INFORMATION PROCESSING SYSTEMS

Overview

- We propose to use meta-learning to effectively incorporate cloth-deformation priors for clothed humans, thus enabling fast fine-tuning (few minutes) for generating neural avatars given only a few monocular depth images of unseen clothed humans and their corresponding SMPL fittings as inputs.

Our Approach

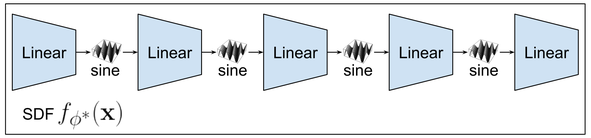
- We follow the same pipeline of recently proposed SCANimate [1], which learns dynamic neural SDFs from dense full-body scans to represent subject/cloth-type specific avatars of clothed humans. Instead of learning subject/cloth-type specific models from scratch, we propose to meta-learn a prior model which can be fast fine-tuned to represent any subject/cloth-type specific neural avatars given only monocular depth frames.



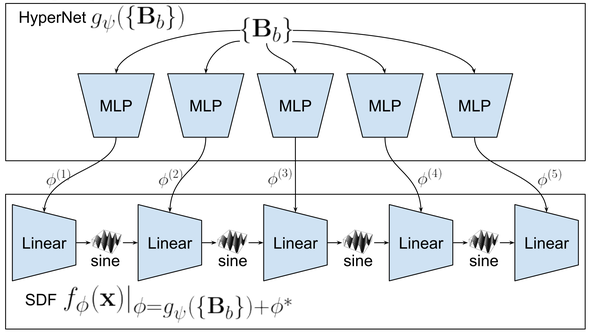
Meta-learned Prior

- Our key contribution is a meta-learned hypernetwork.
- In practice, we decompose the training into two stages 1) meta-learn a static neural SDF 2) meta-learn a hypernetwork which predicts residuals to the parameters of the previously meta-learned static neural SDF.

Training Stage 1: Learn Meta-SDF



Training Stage 2: Learn Meta-Hypernetwork



Experiments

- We quantitatively evaluate ground-truth-to-mesh distances (D_p and D_f) and normal consistency (NC) for pose-interpolation (Int.) on the CAPE dataset. We also conduct large-scale perceptual study (PS) for pose-extrapolation (Ex.).
- Results show that our approach performs well with limited depth input data.

		3D Input			2.5D Input
		NASA	LEAP	SCANimate	Ours
Subj 00122, 00215					
Ex.	PS ↑	0.078	0.314	0.333	**0.5**
Int.	D_p ↓	0.484	0.454	0.586	**0.450**
	D_f ↓	0.327	0.293	0.489	**0.273**
	NC ↑	0.752	0.807	0.793	**0.821**
Subj 00134, 03375					
Ex.	PS ↑	0.182	0.224	0.481	**0.5**
Int.	D_p ↓	0.595	**0.483**	0.629	0.518
	D_f ↓	0.469	**0.340**	0.542	0.367
	NC ↑	0.693	**0.780**	0.755	0.773
Averge per-model training/fine-tuning time (hours)					
		>10	-	>10	1.60

Comparison to baslines

Fine-tune data (%)		100	50	20	10	5	<1
Subj 00122, 00215							
Ex.	PS ↑	0.5	0.471	0.509	0.473	0.373	**0.510**
Int.	D_p ↓	-	**0.450**	0.480	0.512	0.543	0.592
	D_f ↓	-	**0.273**	0.310	0.353	0.391	0.450
	NC ↑	-	**0.821**	0.808	0.795	0.785	0.768
Subj 00134, 03375							
Ex.	PS ↑	**0.5**	0.476	0.424	0.463	0.439	0.387
Int.	D_p ↓	-	**0.518**	0.545	0.576	0.603	0.619
	D_f ↓	-	**0.367**	0.400	0.438	0.471	0.489
	NC ↑	-	**0.773**	0.762	0.753	0.745	0.737
Average per-model training/fine-tuning time (hours)							
		1.60	0.8	0.32	0.16	0.08	0.02

Few-shot learning

[1] Saito et al, SCANimate: Weakly Supervised Learning of Skinned Clothed Avatar Networks. CVPR, 2021.