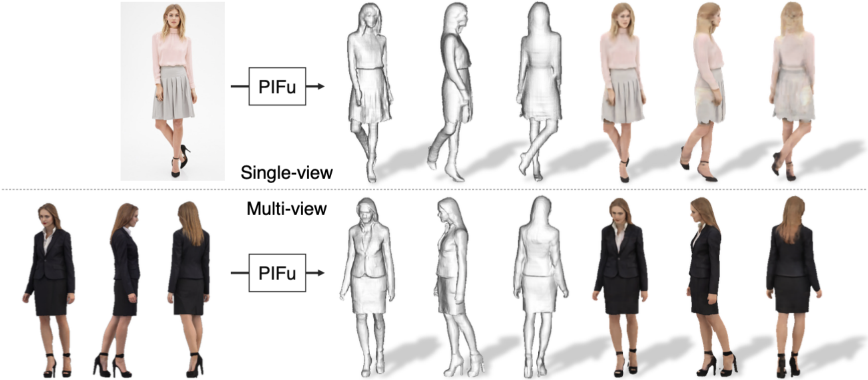
PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization

by Peiyuan Zhu

Match 2020

Problem statement

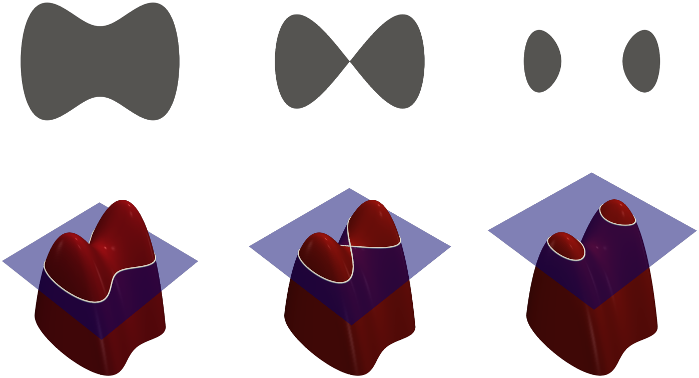


Related work

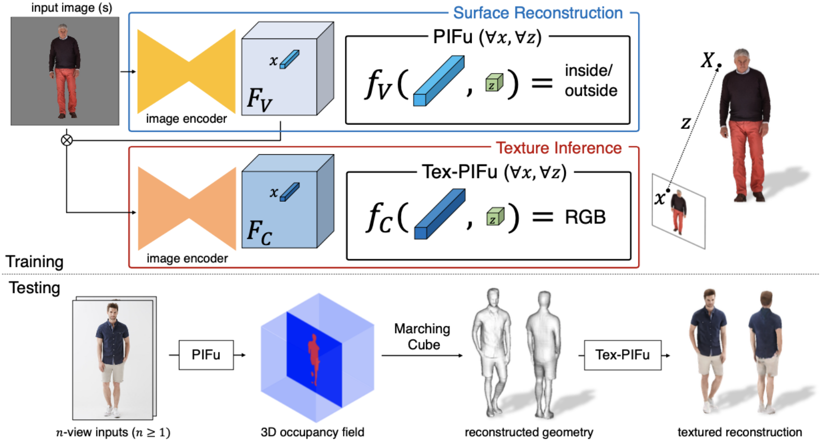
- Other method to solve this problem
    - Voxel representation: memory requirement scales cubically with precision
    - Global representation: cannot capture intricacy of the images
    - Other implicit function method: not spatially aligned
- This paper overcomes these three drawbacks with a simple idea

Idea

Closed surface $\Omega \subset \mathbb{R}^3$ can be represented by implicit function $f(x, y, z) = 0$



Overview


input image (s)
Surface Reconstruction
PIFu (∀x, ∀z)
x
F_V
$f_V($, z $)$ = inside/ outside
image encoder
Texture Inference
Tex-PIFu (∀x, ∀z)
x
F_C
$f_C($, z $)$ = RGB
image encoder
X
z
x
Training
Testing
PIFu
Marching Cube
Tex-PIFu
n-view inputs ($n \geq 1$)
3D occupancy field
reconstructed geometry
textured reconstruction

Single-view reconstruction

- We use single-view reconstruction as an example
- Define
  - f_θ : implicit function parameterized by θ. It takes value of
    - Probability $[0,1]$ of a point is inside the surface
    - RGB color $[1:256]^3$ for texture inference
  - F_η : autoencoder parameterized by η
  - π : projection of 3D point on 2D image
  - z : camera depth of 3-D point
- Implicit function

$$f_\theta\left(F_\eta \circ \pi\left(x\right), z\left(x\right)\right)$$

- Learning f_θ, F_η gives an occupancy/color field
- Occupancy field $\implies$ human geometry by matching cube algorithm
- How to define the optimization procedure?

Single-view reconstruction

- Define distance

$$D\left(f_\theta\left(F_\eta \circ \pi\left(x\right), z\left(x\right)\right) || I\{x \text{ is inside the surface mesh}\}\right)$$

- Use

$$\mathrm{E}_{p(x)}\left|f_\theta\left(F_\eta \circ \pi\left(x\right), z(x)\right) - I\{x \text{ is inside the surface mesh}\}\right|^2$$

- This can be approximated by
  - Sample $x_1, \cdots, x_n \sim p(x)$
  - Calculate $\frac{1}{n}\sum_{i=1}^{n}\left|f_\theta\left(F_\eta \circ \pi\left(x_i\right), z\left(x_i\right)\right) - I\{x_i \text{ is inside the surface}\}\right|^2$
- Use

$$p(x) = \frac{15}{16}N\left(\Omega, 5\right) + \frac{1}{16}\mathrm{Unif}\left(X\right)$$

where Ω is the surface mesh

- Can we use the same architecture for texture inference?

Texture inference

- We can also use this to inference texture
- Simiarly, loss is calculated as

$$\frac{1}{n}\sum_{i=1}^{n}\left|f_\theta\left(F_\eta\left(x_i\right), z(x_i)\right) - \mathrm{ColorOf}\left(x_i\right)\right|$$

- To reduce overfitting, add noise to $x_1, \cdots, x_i$
- How about multiple views?

Multi-view reconstruction

- Suppose there are m views
- First, learn individual latent representations

$$\Phi_i = \tilde{f}_{\theta_i}\left(F_{\eta_i} \circ \pi(x_i), z(x_i)\right)$$

- Second, pool latent features from different views

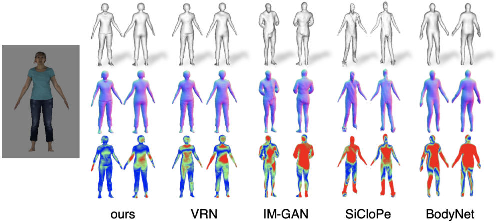
$$\bar{\Phi} = \text{mean}\left(\{\Phi_i\}\right)$$

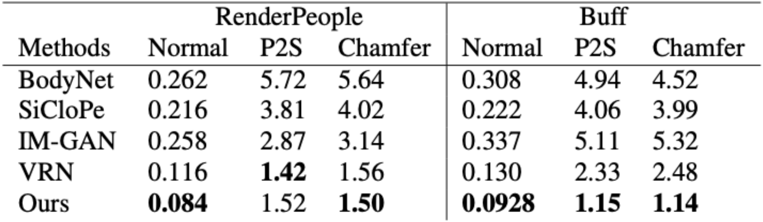
- Feed this pooled feature to a global implicit function

$$f_\lambda\left(\bar{\Phi}\right) = \text{inside/outside, RGB color, etc.}$$

- How does this architecture work on real datasets?

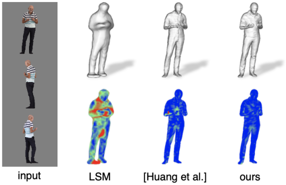
Single-view evaluation

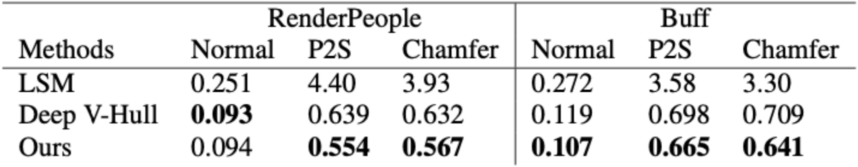




	RenderPeople			Buff		
Methods	Normal	P2S	Chamfer	Normal	P2S	Chamfer
BodyNet	0.262	5.72	5.64	0.308	4.94	4.52
SiCloPe	0.216	3.81	4.02	0.222	4.06	3.99
IM-GAN	0.258	2.87	3.14	0.337	5.11	5.32
VRN	0.116	**1.42**	1.56	0.130	2.33	2.48
Ours	**0.084**	1.52	**1.50**	**0.0928**	**1.15**	**1.14**

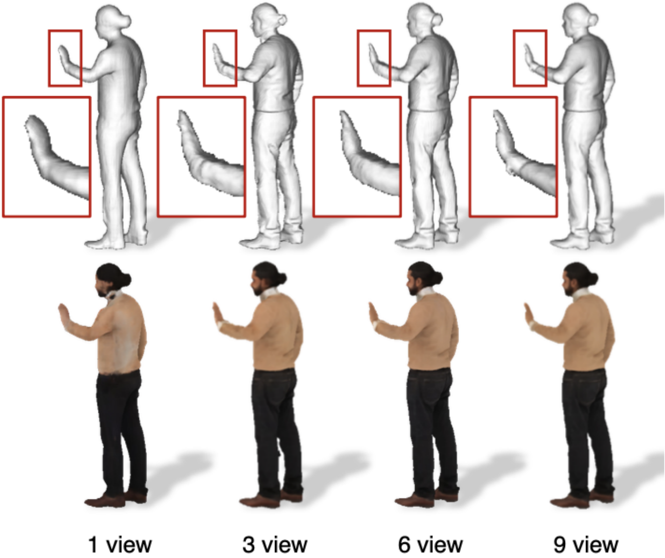
Multie-view evaluation





	RenderPeople			Buff		
Methods	Normal	P2S	Chamfer	Normal	P2S	Chamfer
LSM	0.251	4.40	3.93	0.272	3.58	3.30
Deep V-Hull	**0.093**	0.639	0.632	0.119	0.698	0.709
Ours	0.094	**0.554**	**0.567**	**0.107**	**0.665**	**0.641**

Consistency

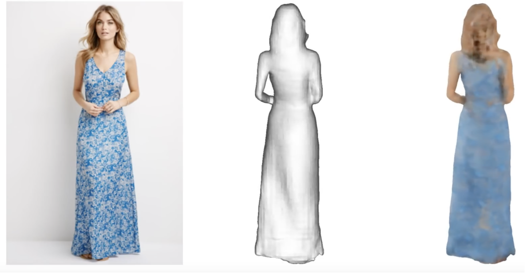


Summary

- Proposed a deep learning framework for human digitization
- Can be generally applied to surface, texture, and multi-view
- Achieved state-of-the-art performance on two benchmark datasets
- Closer to the ground-truth when increase the number of views

Discussion

- From the result, it seems that the extrapolation of the clothes on unseen area of the human were flawed with mixing them up with the human skins and the textures weren't exactly preserved.



- Is it better to perform segmentation first to distinguish human bodies from clothes, then reconstruct them with separate implicit function networks?

Discussion

- From the result of the video, the body shape of a human is far from realistic when the body is behind the human. `https://www.youtube.com/watch?v=S1FpjwKqtPs`
- Can we represent the surface as a differential equation and use neural network to learn the differential operator? When the views are limited, extrapolation by differential equation can be a better way.

The End