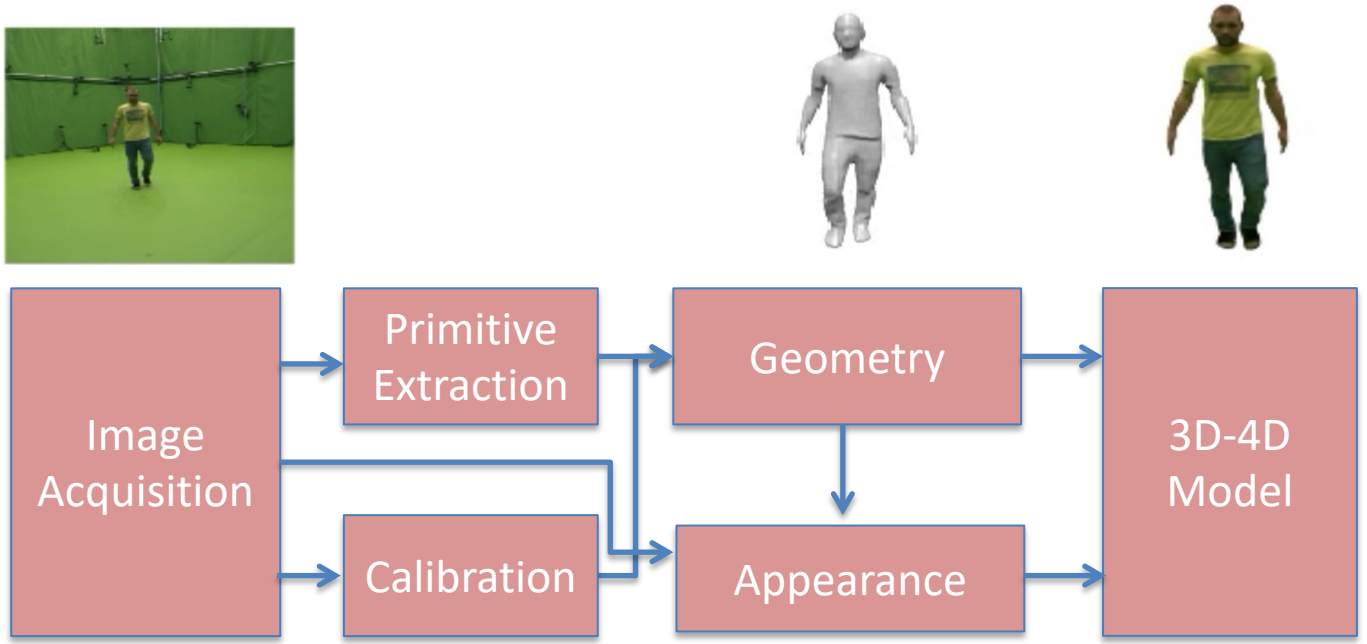
3D Shape Modeling 2

Edmond Boyer
MORPHEO - INRIA Grenoble Rhône-Alpes

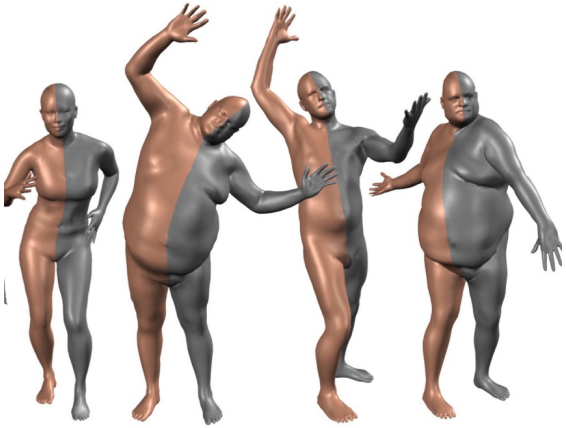
3D Shape Modeling Using Visual Cues



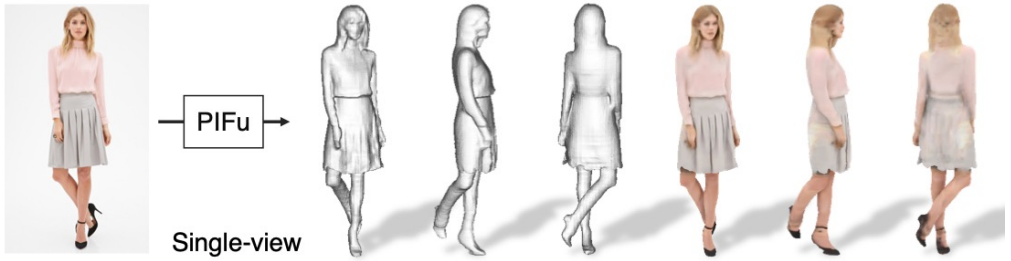
Traditional generative 3D-4D modeling pipeline with no prior model, shape information Y is generated from the image observations X.

Data Driven Approaches

When prior information on the 3D scene to be modeled is available, data driven strategies can be applied. These strategies vary depending on how and where they apply in the inference process, from simple prior statistical models (e.g. human shape models) to full inference with deep neural networks.

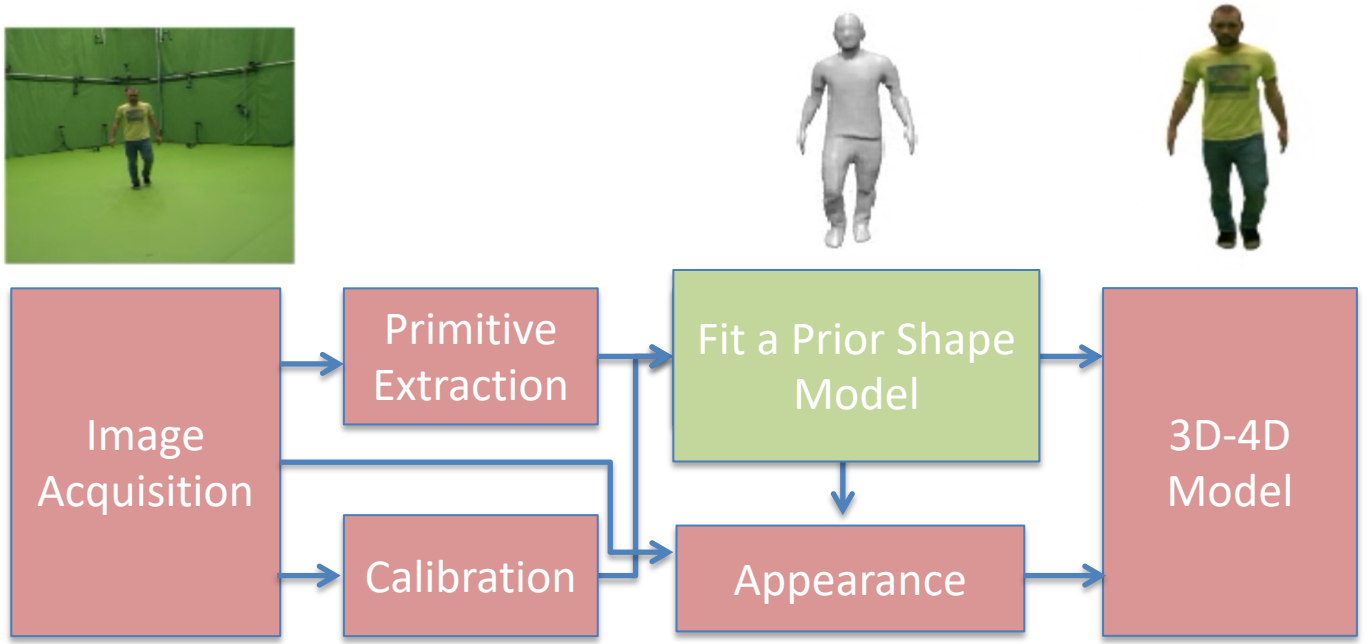


SMPL: A Skinned Multi-Person Linear Model.
M. Loper et al., MPI Saarbrücken, ACM ToG'15.



Pifu:Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization
S. Saito et al. ICCV 2019.

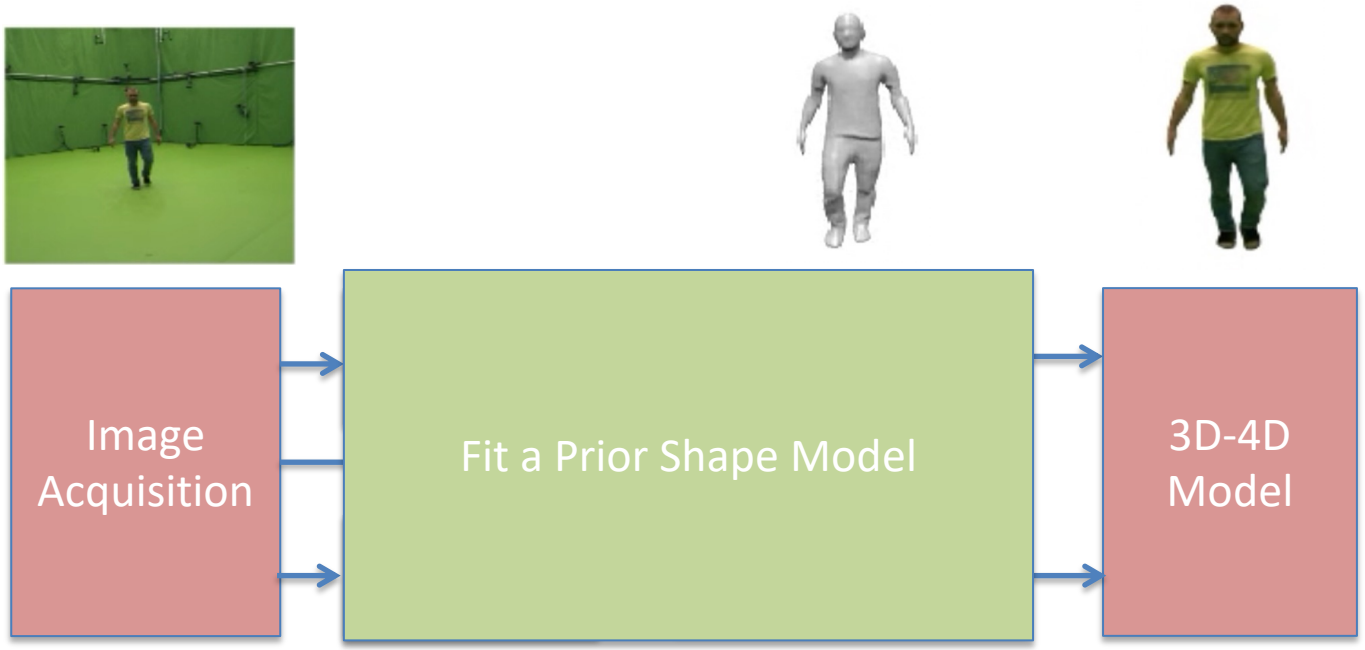
3D Shape Modeling Using Visual Cues



Prior shape model: Shape information is generated by fitting a (parametric) shape model Y(p) to the observations X:

$p^* = \mathrm{Min}_p \; | X - F(Y(p)) |$ with F() maps Y to X (e.g. 3D-2D projections)

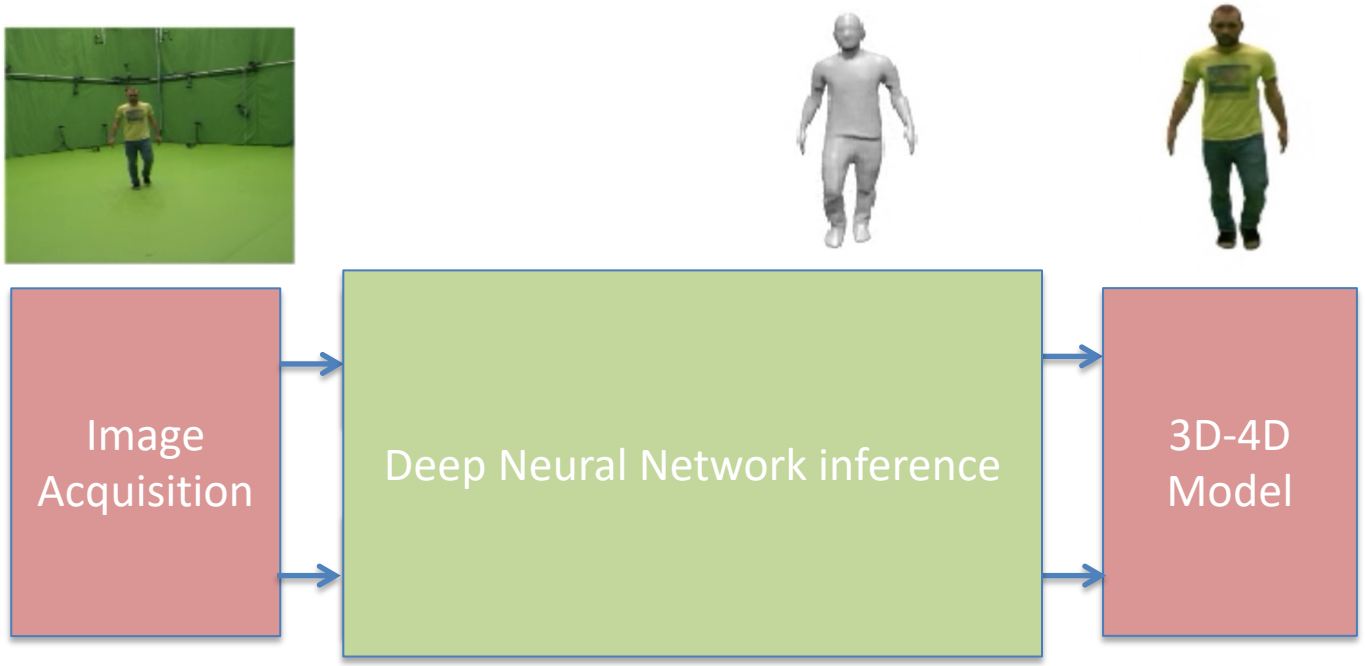
3D Shape Modeling Using Visual Cues



Prior shape model: Shape information is generated by fitting a (parametric) shape model Y(p) to the image observations X:

$$p^* = \mathrm{Min}_p \; | X - F(Y(p)) |$$

3D Shape Modeling Using Visual Cues

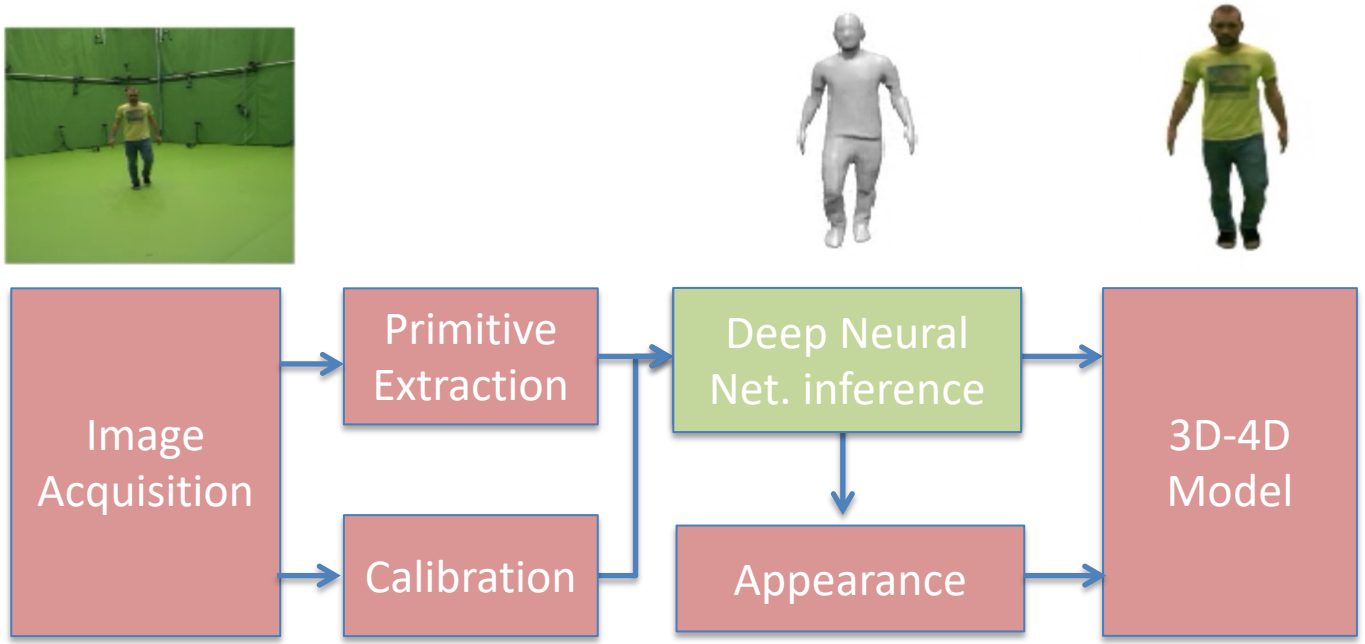


Direct Inference: Shape information X is inferred from the observations Y using a learning based approach, e.g. CNN:

$X \rightarrow Y_X$ where $\rightarrow$ is a network trained to minimize:

$\sum_i | Y_i - Y_{xi} |$ over a dataset of known pairs (Y_i, X_i).

3D Shape Modeling Using Visual Cues



Direct Inference: Shape information X is inferred from the observations Y using a learning based approach, e.g. CNN:

$X \rightarrow Y_X$ where $\rightarrow$ is a network trained to minimize:

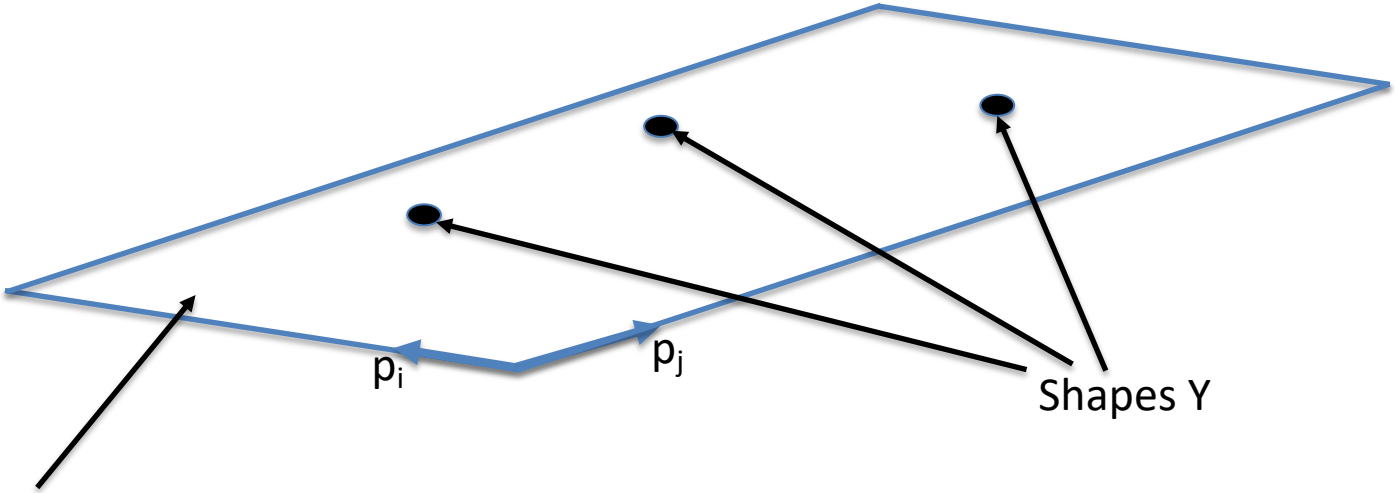
$\sum_i | Y_i - Y_{xi} |$ over a dataset of known pairs (Y_i, X_i).

Prior Models

Dyna: A Model of Dynamic Human Shape in Motion
Pons-Moll, Gerard and Romero, Javier and Mahmood, Naureen and Black, Michael
Siggraph 2015

Prior Models

Shape Spaces



Parametric shape space with n parameters $Y(p_{[1..n]})$

Prior Models

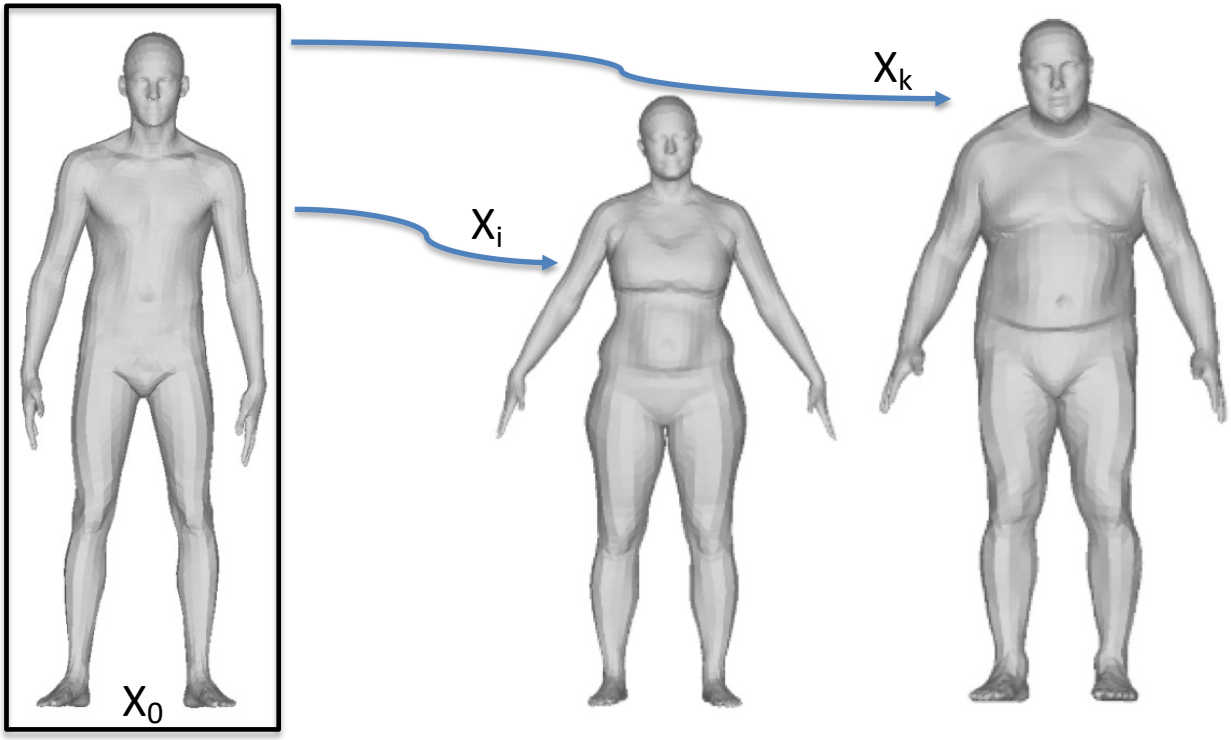
Shape Spaces with PCA

[Caesar human dataset, Siggraph 2003, Allen, Curless, Popovic]

PCA strategy allows to represent shapes with « eigen » shapes

Prior Models

Shape Spaces with PCA



Assume that we have a shape template X_0 fitted to n shapes X_i with different properties, e.g biometric properties or poses.

Prior Models

Shape Spaces with PCA, Intuition:

The template X_0 is composed of m vertices with 3 coordinates (x,y,z):

$$X_0 = [x_{01}, y_{01}, z_{01}, \ldots, x_{0m}, y_{0m}, z_{0m}]^t$$

and we have for each registered shape the corresponding vector X_i of vertex locations.

Centering all shapes, $\overline{X}_i = X_i - \sum_i X_i/n$, and stacking the coordinate vectors together we get a n×3m matrix:

$$M = \begin{bmatrix} \overline{x}_{11} & \cdots & \overline{z}_{1m} \\ \vdots & \ddots & \vdots \\ \overline{x}_{n1} & \cdots & \overline{z}_{nm} \end{bmatrix}$$

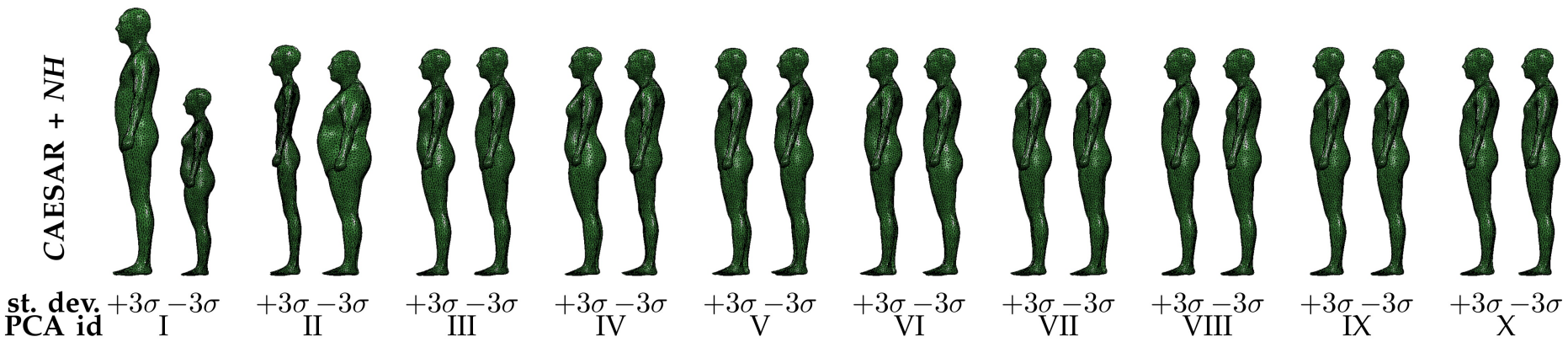
The 3m×3m covariance matrix writes: $C = 1/n\ M^t \cdot M$, and encodes variations from the average shape over the n shapes.

The 3m eigenvectors E_i of the matrix M form then an orthonormal basis of the shape space spanned by the instances of X_0. They capture the main shape variations (over the n input shapes) in decreasing order with their respective eigenvalues and any instance X of the template X_0 writes as a linear combination: $X = \sum_i \alpha_i E_i$.

In practice PCA captures rather well human shape variations but not well human pose variations and usually combined strategies are used (PCA shape transformation + skeleton driven pose transformation) as in the Scape and SMPL models.

Prior Models

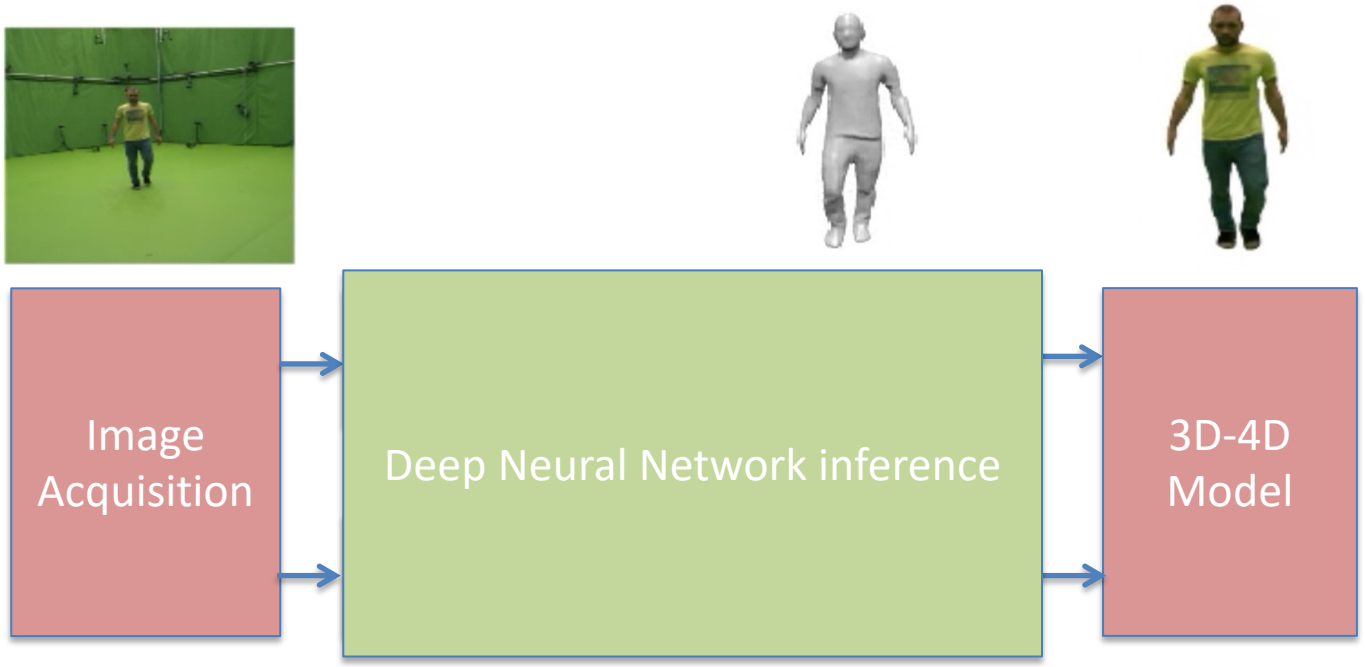
Shape Spaces: PCA illustration



MPII Human Shape Pishchulin, Wuhrer, Helten, Theobalt, Schiele

Variations along the first 10 PCA components

Deep Learning Strategies



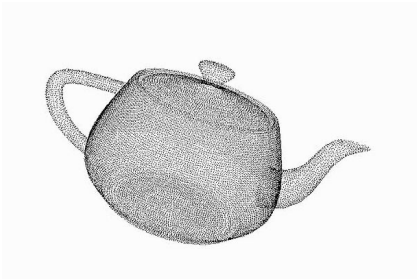
Direct Inference: Shape information X is inferred from the observations Y using a learning based approach, e.g. CNN:

$X \rightarrow Y_X$ where $\rightarrow$ is a network trained to minimize:

$\sum_i | Y_i - Y_{xi} |$ over a dataset of known pairs (Y_i, X_i).

Deep Learning Strategies

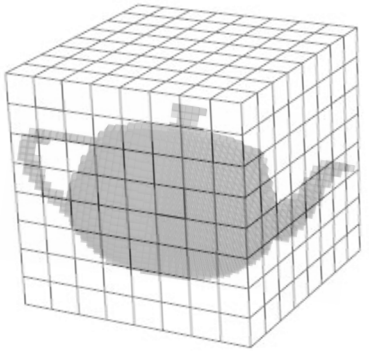
Deep learning strategies replace part of, or the entire, modeling pipeline with a data driven inference approach trained on examples. Usually considering pixel information directly, they can output different information such as: point clouds, meshes, voxels or more intermediate information such as photoconsistency.



Point clouds

3D Meshes

Explicit (Lagragian) representations

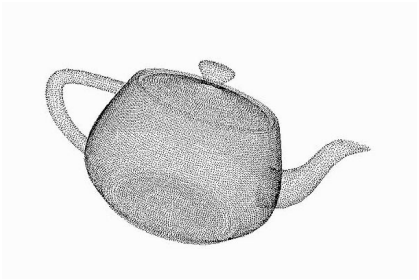


Voxels

Implicit (Eulerian) representations

Deep Learning Strategies

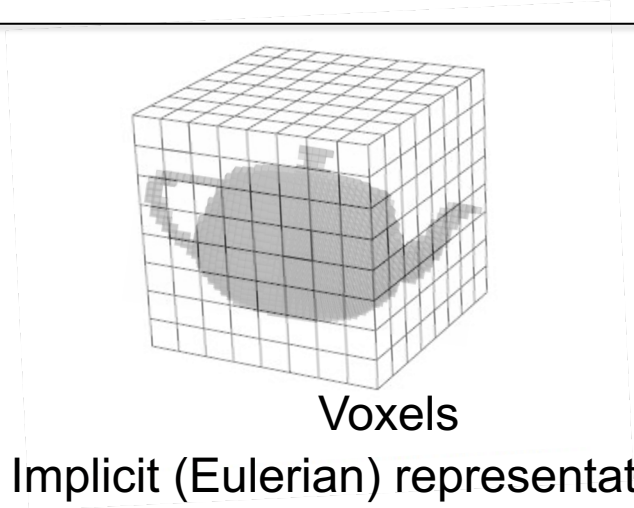
Deep learning strategies replace part of, or the entire, modeling pipeline with a data driven inference approach trained on examples. Usually considering pixel information directly, they can output different information such as: point clouds, meshes, voxels or more intermediate information such as photoconsistency.



Point clouds

3D Meshes

Explicit (Lagragian) representations



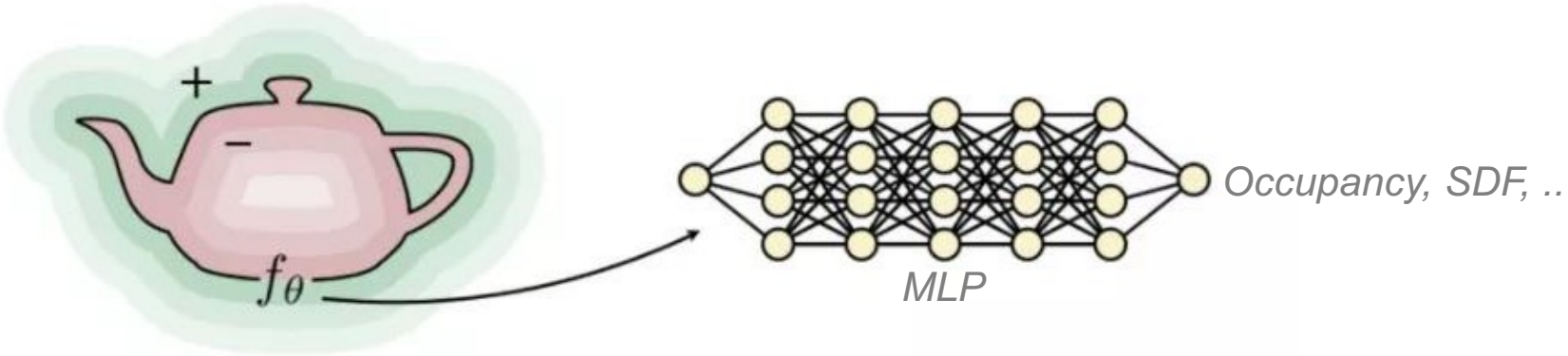
Voxels

Implicit (Eulerian) representations

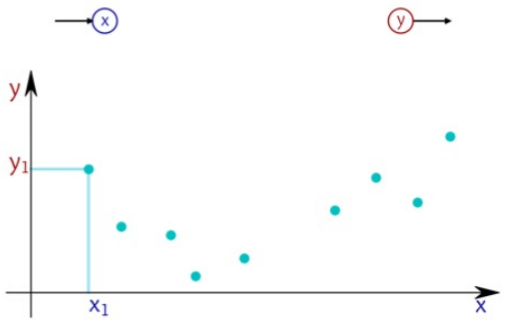
Discrete representations !!

Deep Learning Strategies

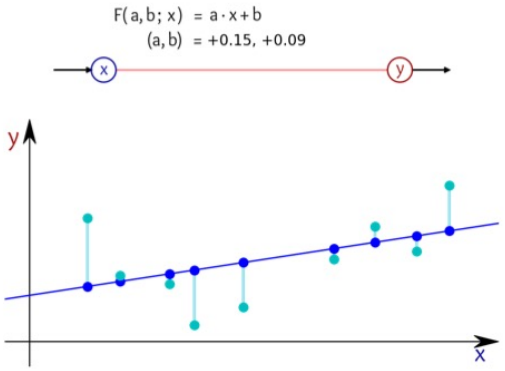
In practice it seems easier to predict the occupancy in 3D (classification problem) and there is a strong new trend towards the inference of implicit shape representations through MLPs (fully connected layers).



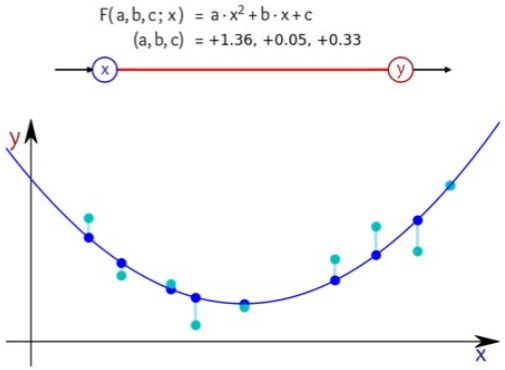
+ Continuous shape representations
- Implicit to explicit tool required
- Rather local strategy



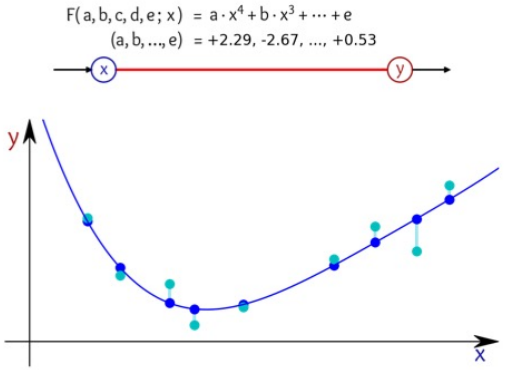
(a) Dataset.



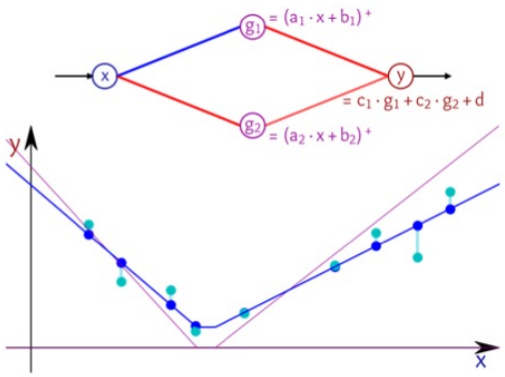
(b) Linear model.



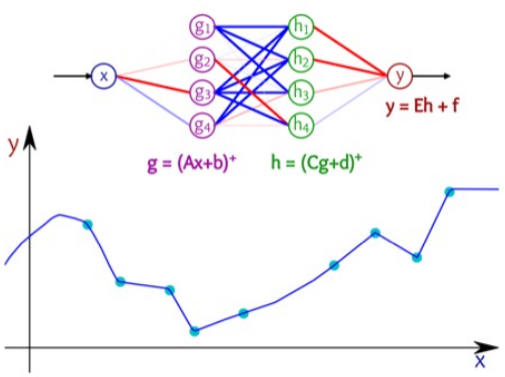
(c) Quadratic model.



(d) Quartic model.



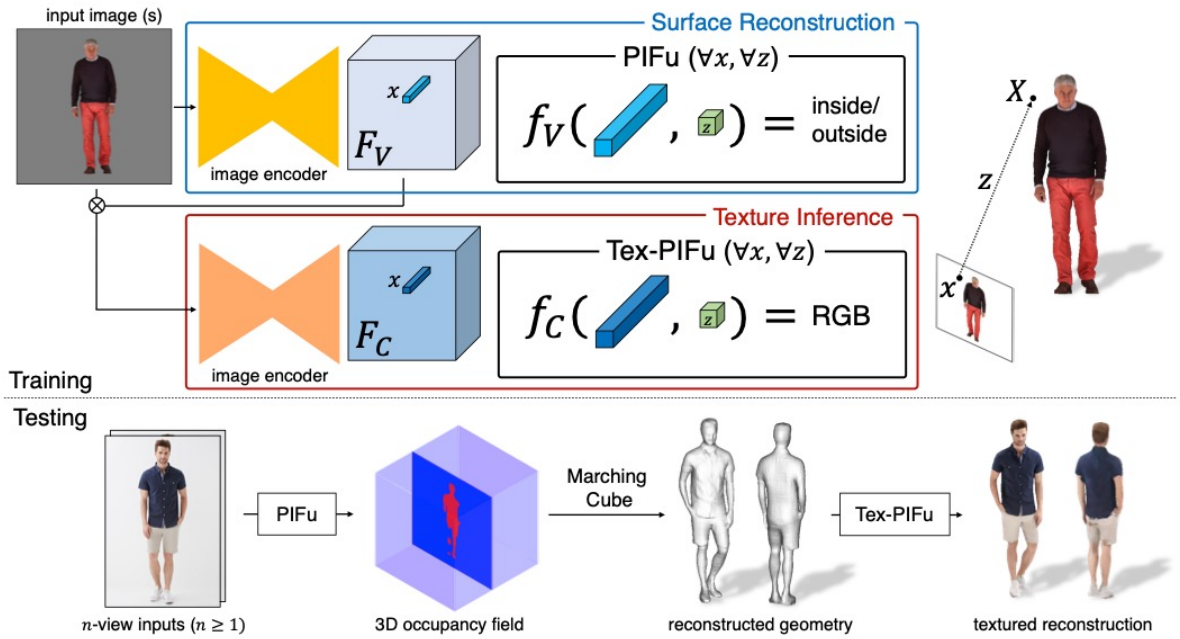
(e) Two hidden variables.



(f) Multi-layer perceptron.

Deep Learning Strategies

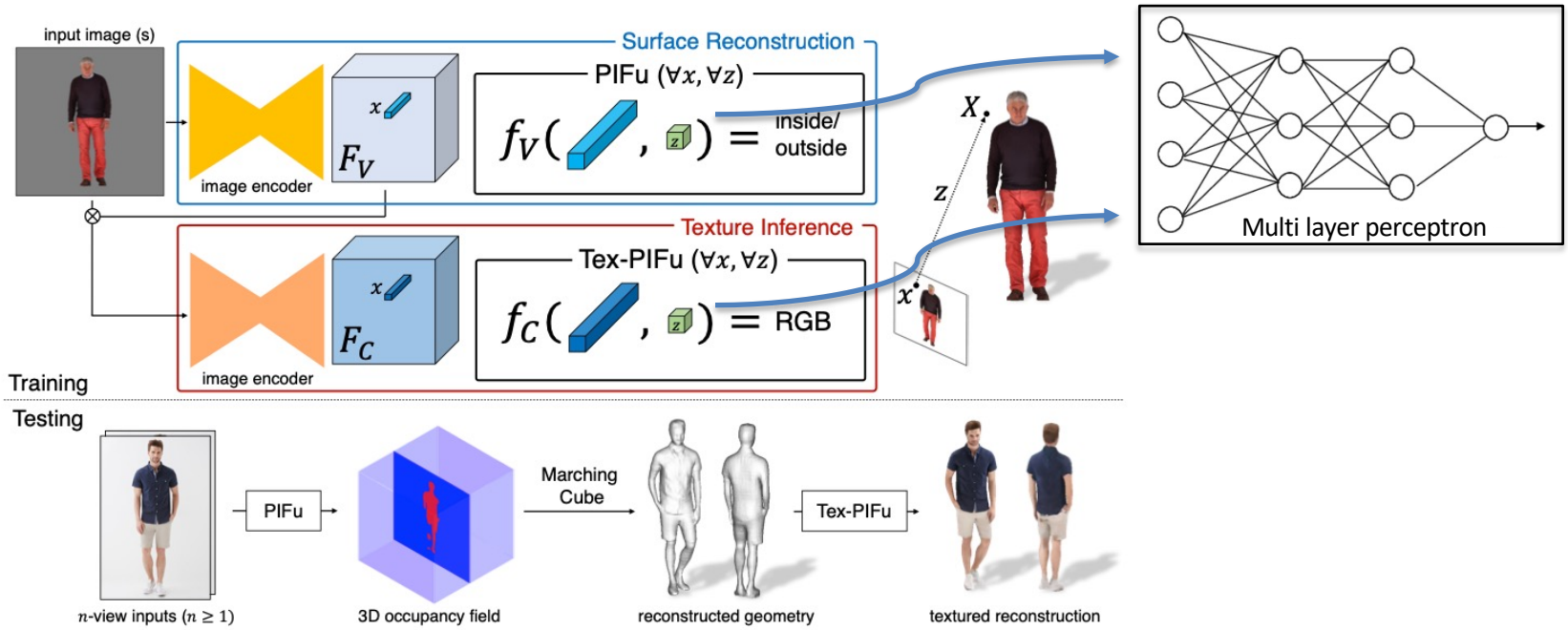
Illustration with The PiFu Approach.



PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization
Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, Hao Li

Deep Learning Strategies

Illustration with The PiFu Approach.

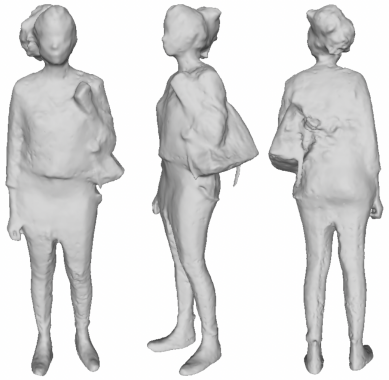


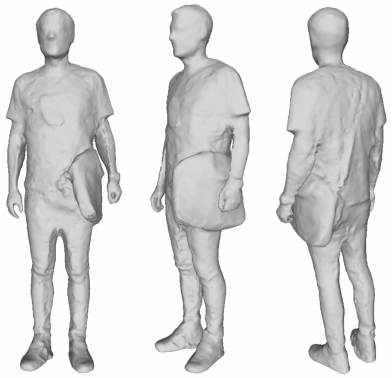
PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization
Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, Hao Li

Deep Learning Strategies

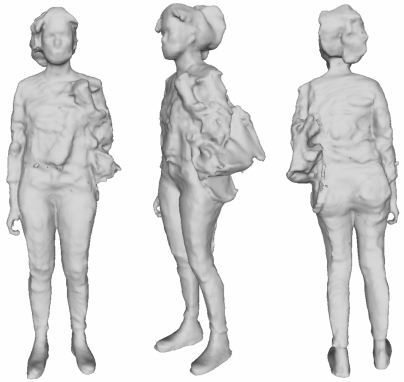
Illustration with The PiFu Approach.

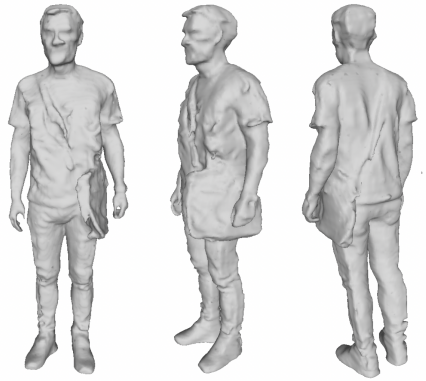
Deep Learning Strategies





Multi-View stereo (64 views)

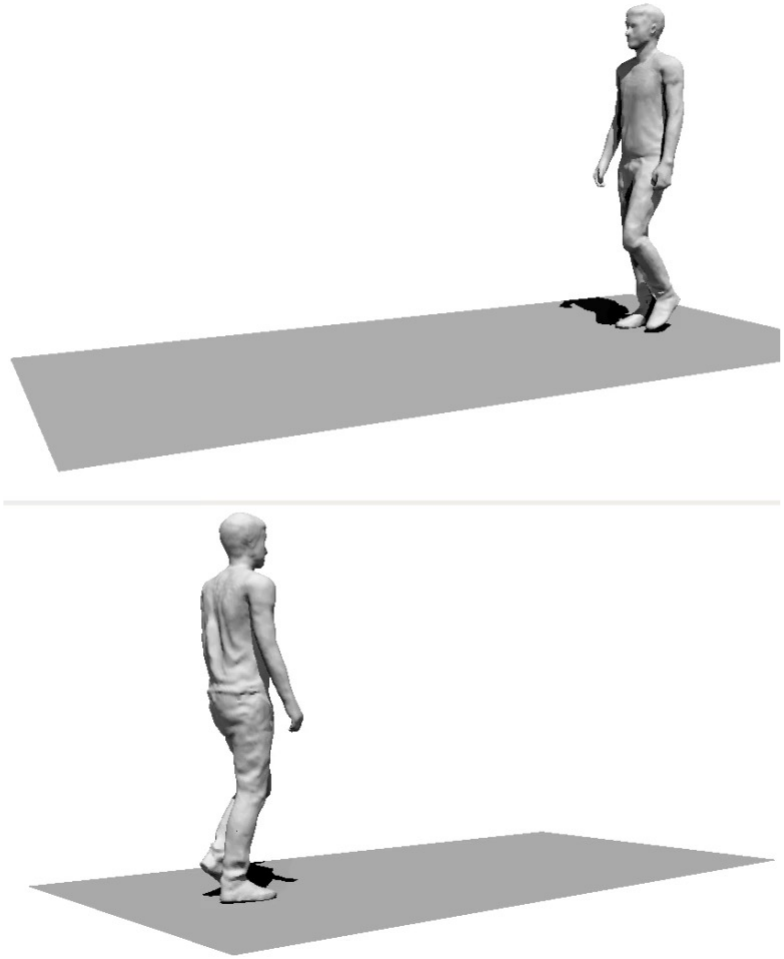




4 view neural implicit modeling

Implicit 3D Representations of Dressed Humans from Sparse Views
[P. Zins, S. Wuhrer, T. Tung, E. Boyer, 3DV 2021]

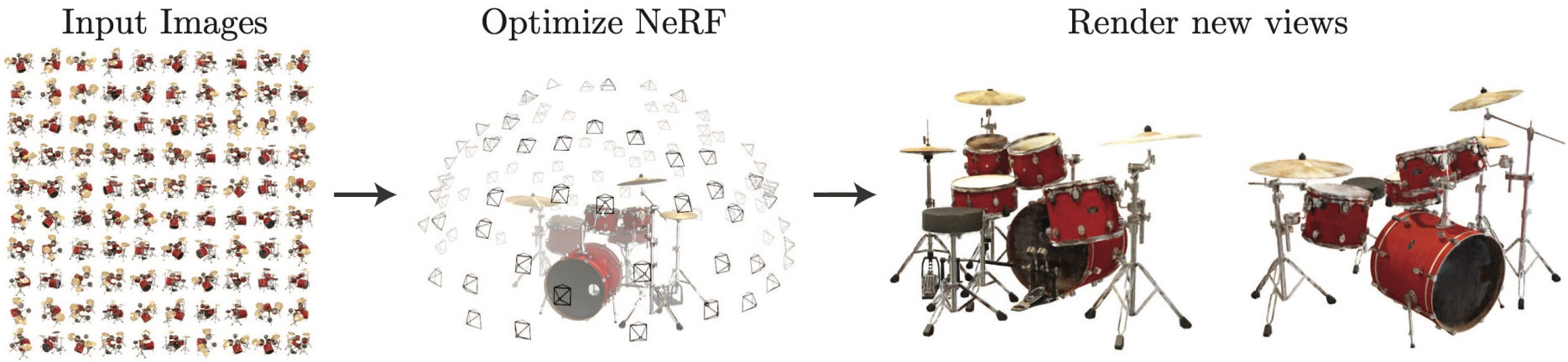
Deep Learning Strategies



4 view neural implicit modeling

Deep Learning Strategies

Another illustration of this new trend goes even further and replace the traditional 3D shape represention: geometry + appearance with a representation encoded by a network.

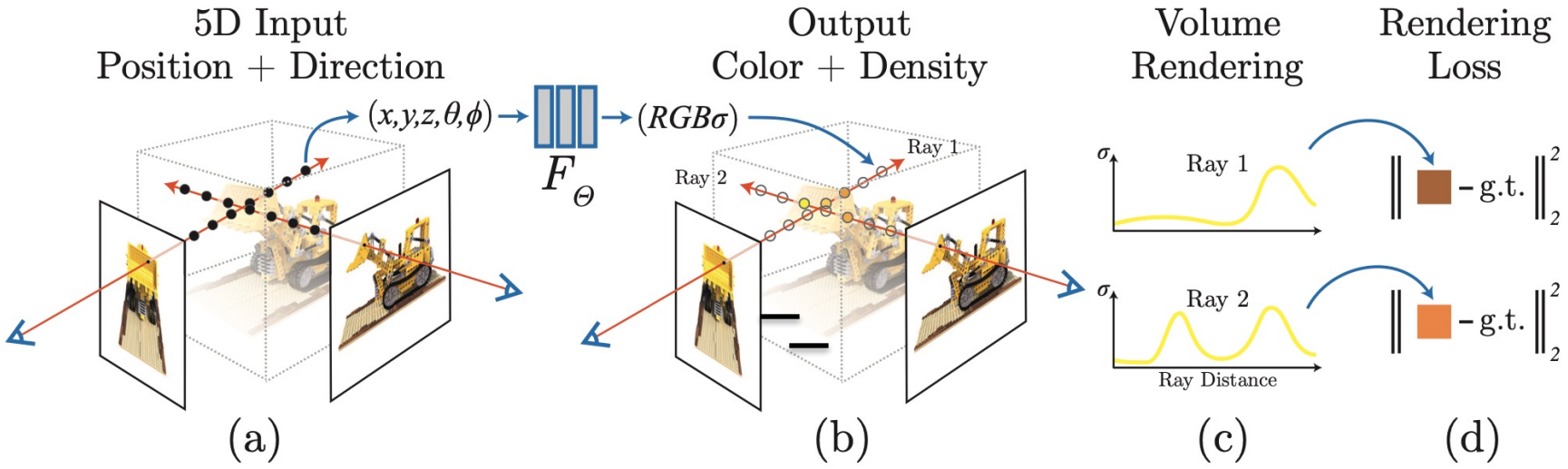


NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis
Ben Mildenhall Pratul P. Srinivasan* Matthew Tancik Jonathan T. Barron Ravi Ramamoorthi Ren Ng
ECCV 2020

Deep Learning Strategies

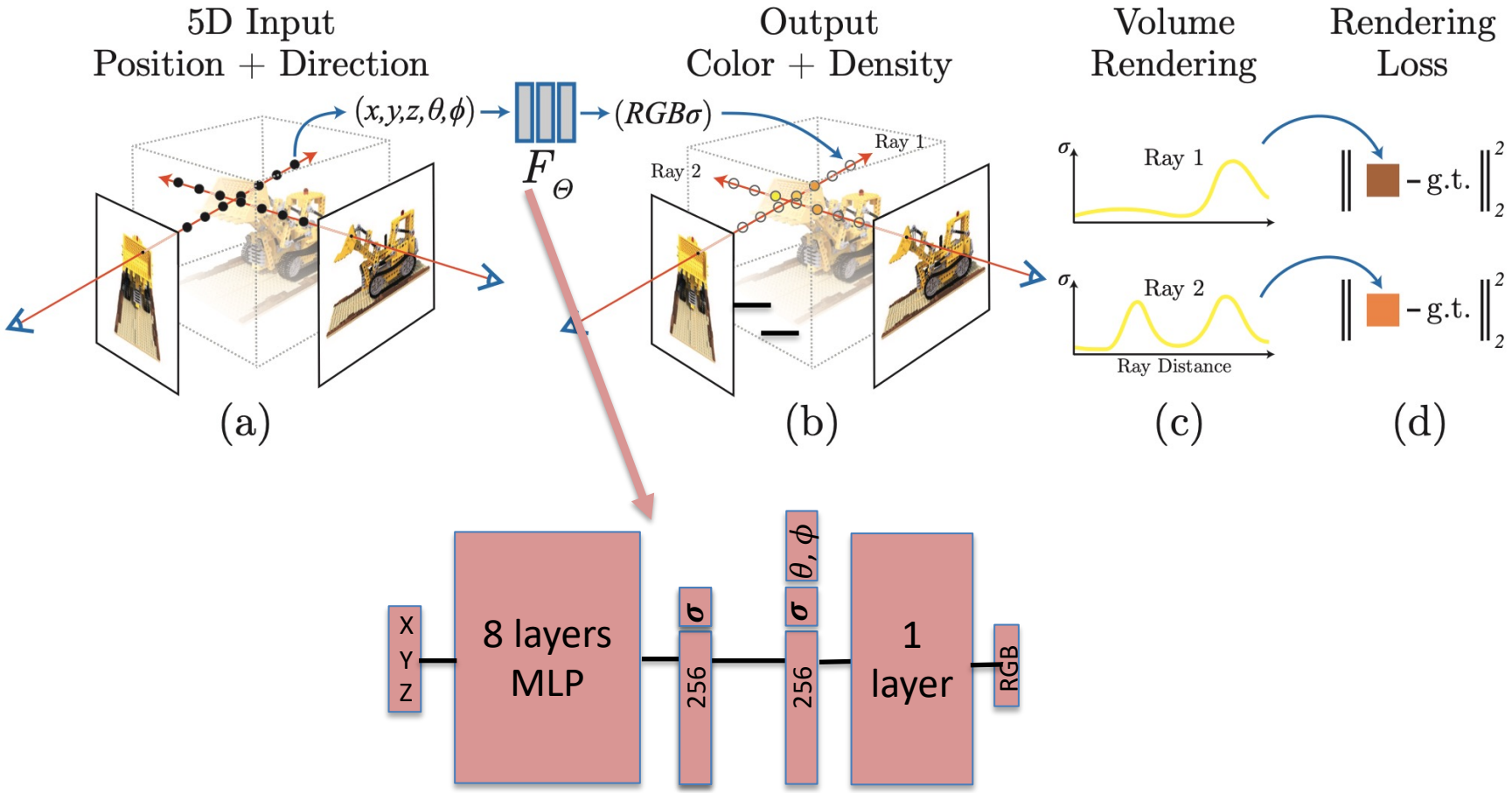
NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis
Ben Mildenhall Pratul P. Srinivasan* Matthew Tancik Jonathan T. Barron Ravi Ramamoorthi Ren Ng
ECCV 2020

Deep Learning Strategies



NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis
Ben Mildenhall Pratul P. Srinivasan* Matthew Tancik Jonathan T. Barron Ravi Ramamoorthi Ren Ng
ECCV 2020

Deep Learning Strategies



Data Driven 3D Vision

NeRF++: Analyzing and Improving Neural Radiance Fields,
Kai Zhang, Gernot Riegler, Noah Snavely, Vladlen Koltun,

➡ The radiance field is ambiguous in Nerf -> the unit sphere can explain the input images but will generate incorrect new images.

➡ Nerf++ conjectures that it works anyway with Nerf for 2 reasons:

1. Incorrect geometries force regularization.
2. The scene location and the viewing direction are treated asymetrically. Introducing the direction later in the network favors smooth surface reflectance (RGB) with fewer network parameters to explain them.

Data Driven 3D Vision

Critical aspects with MLP implicit modeling:

- (MLPs) Lack the capacity to model high order signal information:
  - Due to piecewise linear approximation.
  - Due also to localized (dirac like) losses ?

- Compensate with data representation (higher order information encoded, e.g. NERF)
- Compensate with activation function (less linear space model)
- Another issue is 3D Sampling when training

Data Driven 3D Vision

Data representation

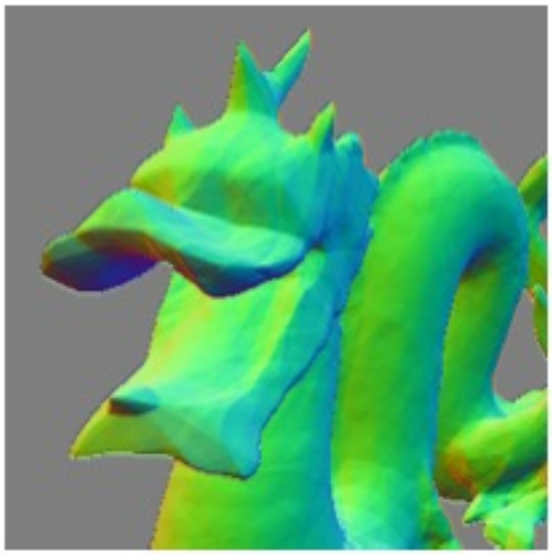
[Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains, Tancik, Srinivasan, Mildenhall, Fridovich-Keil, Nithin Raghavan, Singhal, T. Barron, Ng - NeurIPS 2020]

$$\gamma(\mathbf{v}) = \left[a_1 \cos(2\pi \mathbf{b}_1^{\mathrm{T}} \mathbf{v}), a_1 \sin(2\pi \mathbf{b}_1^{\mathrm{T}} \mathbf{v}), \ldots, a_m \cos(2\pi \mathbf{b}_m^{\mathrm{T}} \mathbf{v}), a_m \sin(2\pi \mathbf{b}_m^{\mathrm{T}} \mathbf{v})\right]^{\mathrm{T}}$$

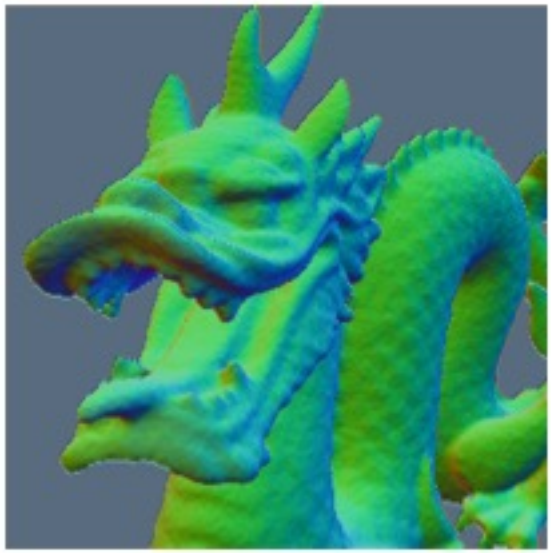
setting $a_j = 1$ and randomly sampling $\mathbf{b}_j$ from an isotropic distribution

3D shape regression

$(x, y, z) \rightarrow$ occupancy



(x,y,z)



γ(x,y,z)

Data Driven 3D Vision

Data representation

Positional encoding [NERF] : $\gamma : \mathbb{R} \rightarrow \mathbb{R}^{2L}$

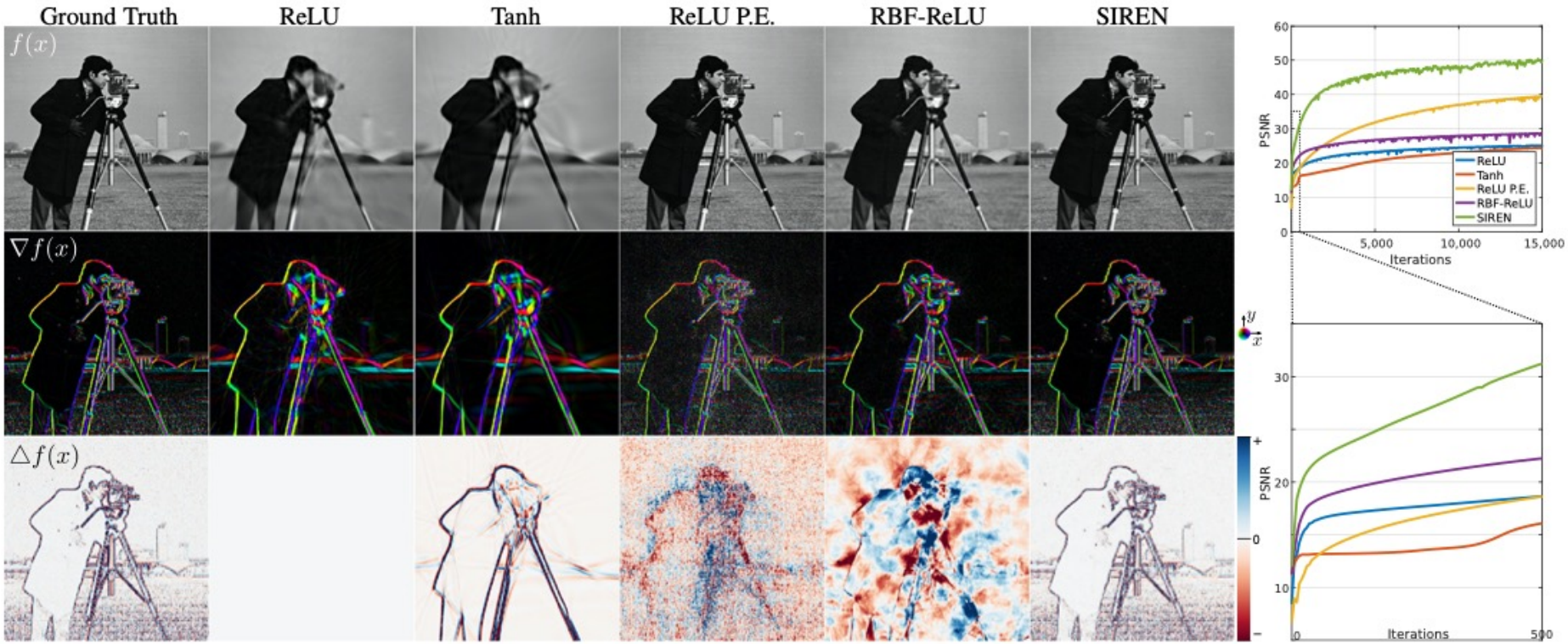
$$\gamma(p) = \left(\sin\left(2^0\pi p\right), \cos\left(2^0\pi p\right), \cdots, \sin\left(2^{L-1}\pi p\right), \cos\left(2^{L-1}\pi p\right)\right).$$

Ground Truth | No Positional Encoding | Complete Model

Data Driven 3D Vision

Activation function: SIREN (sinusoidal representation networks)

[Implicit Neural Representations with Periodic Activation Functions
Sitzmann, Martel, Bergman, Lindell, Wetzstein - NeurIPS2020]



MLP: (x,y) -> (R,G,B) trained on a single image with different activation functions

Data Driven 3D Vision

Another experiment with MLP based 3D Modeling: Signed distance function regression: [ON THE EFFECTIVENESS OF WEIGHT-ENCODED NEURAL IMPLICIT 3D SHAPES Davies, Nowrouzezahrai, Jacobson – submitted to ICLR 2021]

A MLP with 8 layers of hidden size 32 (and ReLU activations) encodes the SDF for a single model.

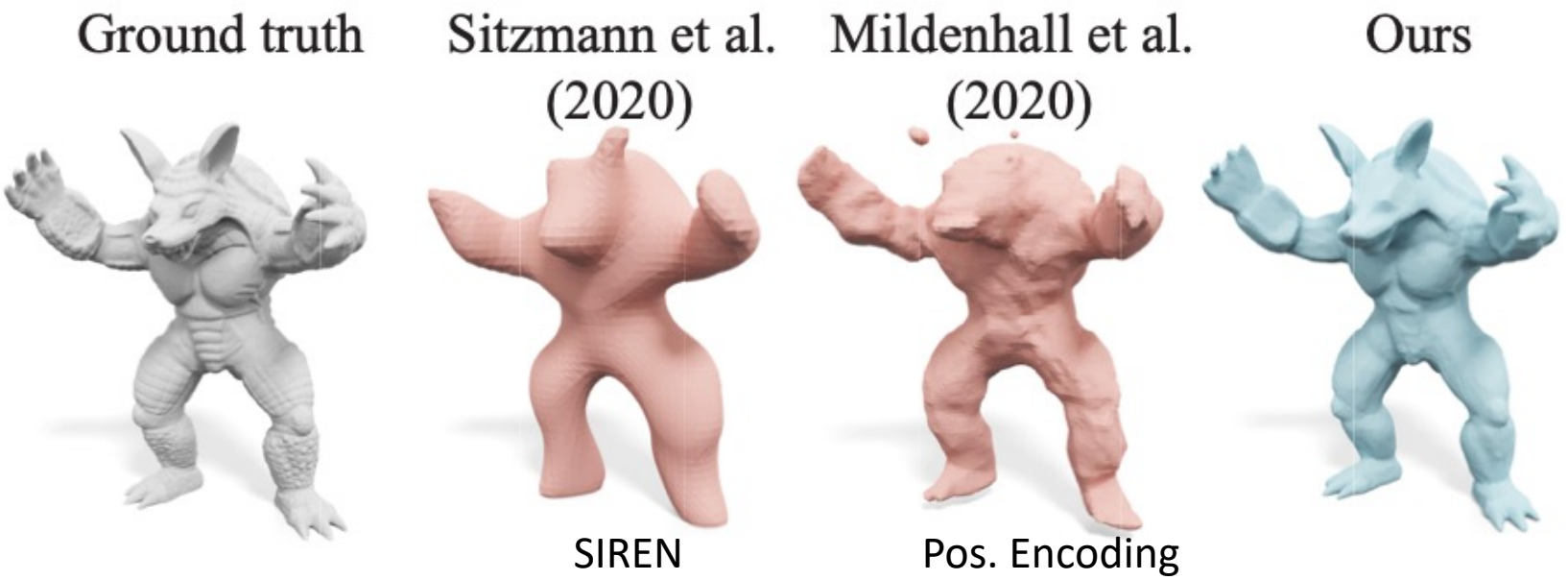
Importance sampling is used at training time:

1. Uniformly sample 10M points in a bounding volume.
2. Subsample 1M points with respect to surface distance probabilities

Data Driven 3D Vision

Signed distance function regression:

[ON THE EFFECTIVENESS OF WEIGHT-ENCODED NEURAL IMPLICIT 3D SHAPES
Davies, Nowrouzezahrai, Jacobson – submitted to ICLR 2021]



Authors mention that with hidden size 64, SIREN starts to give better results than ReLU

Conclusion

- Implicit neural modeling is efficient but investigation still required to better formalize.
- Data driven strategies are a game changer in 3D modeling.