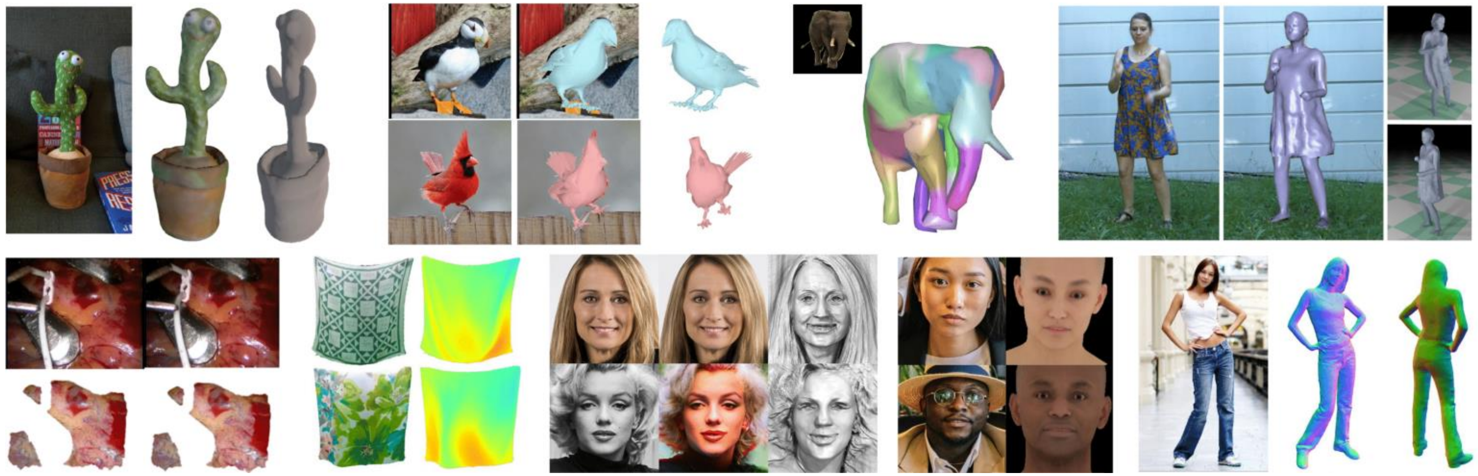
State of the Art in Dense Monocular Non-Rigid 3D Reconstruction



Edith Tretschk* Navami Kairanda* Mallikarjun B R Rishabh Dabral Adam Kortylewski
Bernhard Egger Marc Habermann Pascal Fua Christian Theobalt Vladislav Golyanik

max planck institut informatik
UNI FREIBURG
FAU Friedrich-Alexander-Universität Erlangen-Nürnberg
EPFL

Motivation

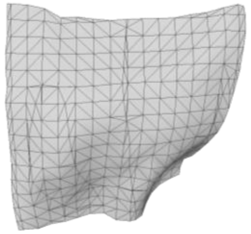
- Goal: 3D reconstruction of dynamic objects from monocular images

Monocular 2D input

?

→ 3D

Lots of ambiguity:
- Depth?
- Occlusions?
- Texture vs. wrinkles?
- Texture vs. illumination?
- Correspondences?
- …



Reconstructed 3D geometry

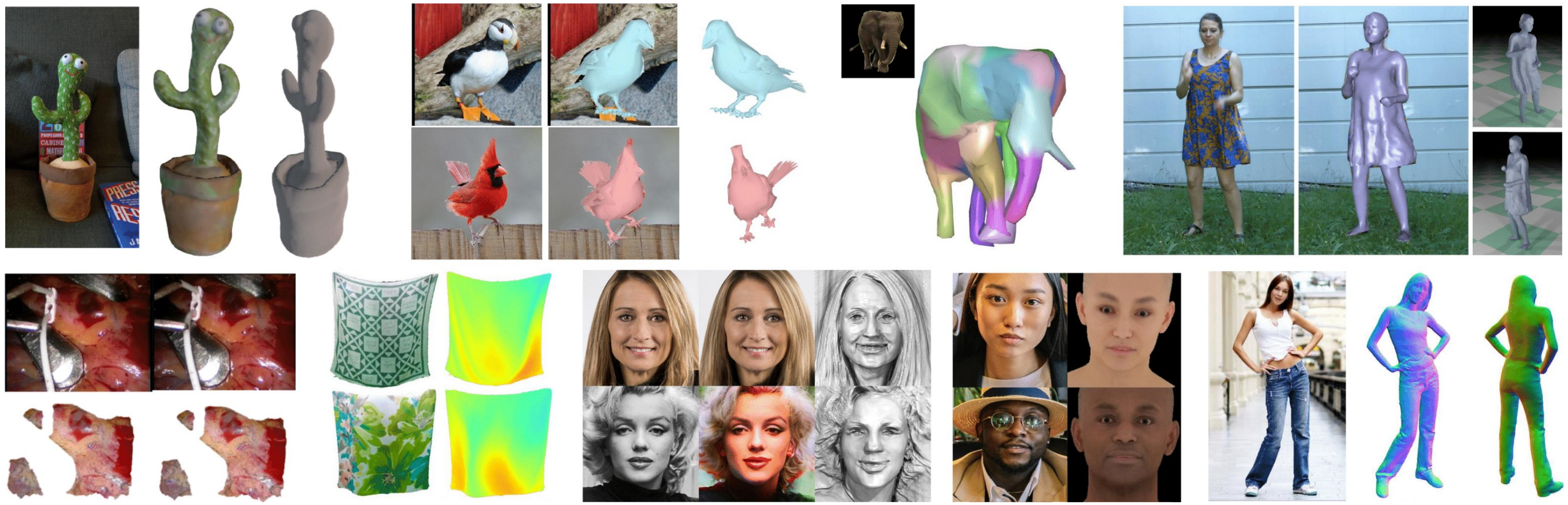
Reconstructed 3D geometry with texture

Kairanda *et al.* 2022

- Thank you to the authors of all the works in this STAR!
  - We tried our best and apologize for any mistakes we made!

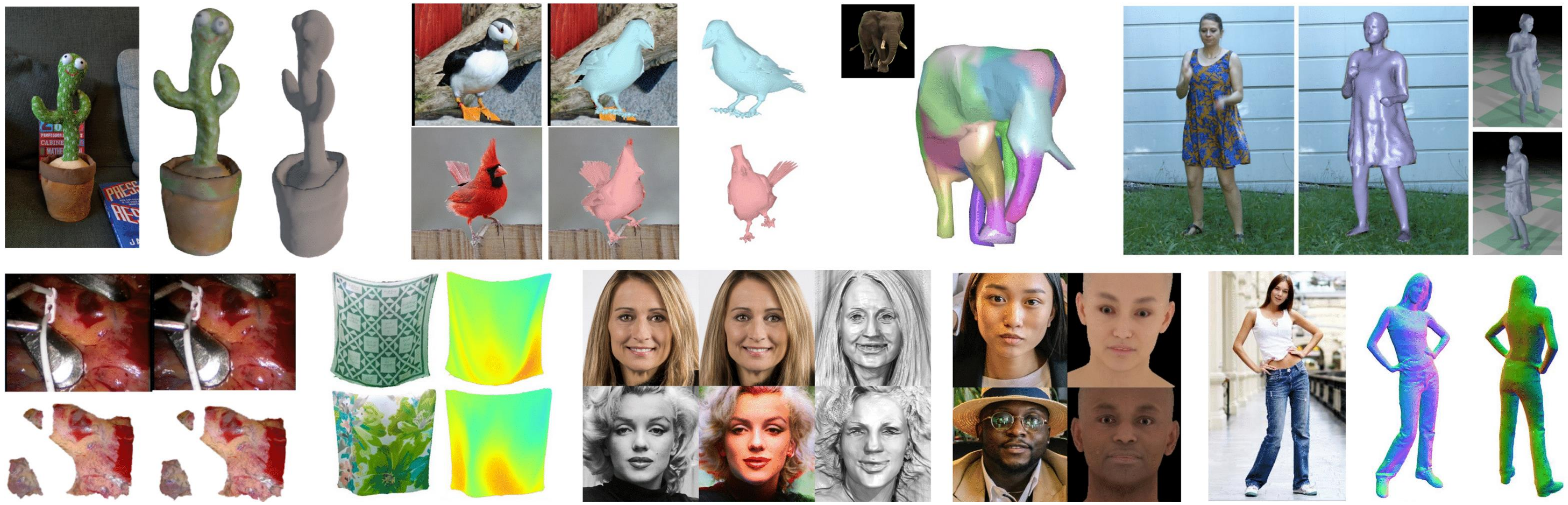
Motivation

- Goal: 3D reconstruction of dynamic objects from monocular images
- Why? Make real world accessible to downstream tasks:
  - Novel view synthesis (telepresence, virtual reality)
  - Geometry acquisition for scene modelling (robotics, augmented reality)
  - Virtual asset creation (video games)
  - Editing for visual content creation (VFX, social media)
  - Motion analysis (physics, biology)



Motivation

- Goal: 3D reconstruction of dynamic objects from monocular images
- Why this STAR now? Several breakthroughs in recent years:
  - Parametric models
  - Neural scene representations
  - Deep learning
  - High-quality, large-scale datasets
  - Powerful hardware



Scope: Dense Monocular Non-Rigid 3D Reconstruction

Dense	Monocular	Non-Rigid	3D	Reconstruction
• Entire surface, not just sparse keypoints	• Single RGB camera • No active sensor (like depth cameras) • Why? Easily accessible to everyone, no specialized setup and synchronization	• Only deformable objects, not static • But: Exclude human-specific methods that *only* estimate parameters of statistical shape models	• True 3D representation • Not image-based or intermediate (like light fields)	• Represent observed state of the scene • Does not need to be generative or editable

Two typical cases:

- Video: Single video of one scene
  → Temporal information
- Image collection: Many images, each of a different scene
  → No temporal information

Structure

1. Introduction
2. Fundamentals
3. State-of-the-Art Methods
    1. General Objects
        1. Shape from Template
        2. Non-Rigid Structure from Motion
        3. Neural Scene Representations
        4. Others
        5. Learned Prior
    2. Humans
    3. Faces
    4. Hands
    5. Animals
4. Emerging Areas
    1. Physics
    2. Event Cameras
5. Open Challenges
6. Social Implications
7. Conclusion

2

Fundamentals

Overview

- Representing deformations
- Rendering and data terms
- Challenges and priors to tackle them

Geometry ~~Representations~~ Functions

- We split “representation” into its two components:
  - Function: Input-output relation
  - Parametrization: How to actually compute the function
- Typical geometry functions to represent a surface $S \subset \mathbb{R}^3$:
  - Indicator function:

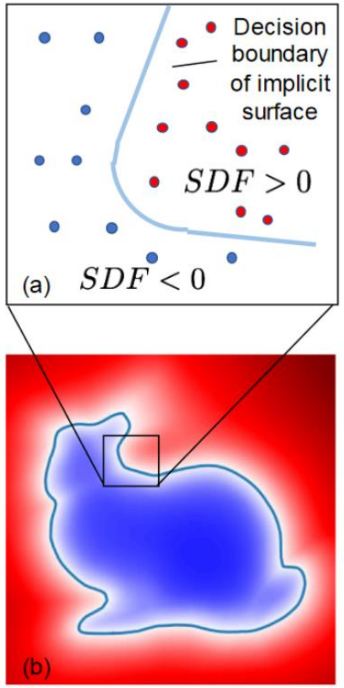
$$s(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in S \\ 0 & \text{else} \end{cases}$$

  - A level-set function:

$$s(\mathbf{x}) = \min_{\mathbf{y} \in S} ||\mathbf{x} - \mathbf{y}||_2$$

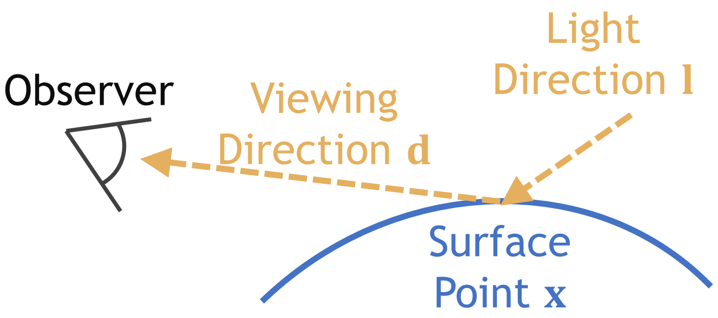
  - A density function:

$$v(\mathbf{x}) = \text{density}(\mathbf{x})$$



Park *et al.* 2019

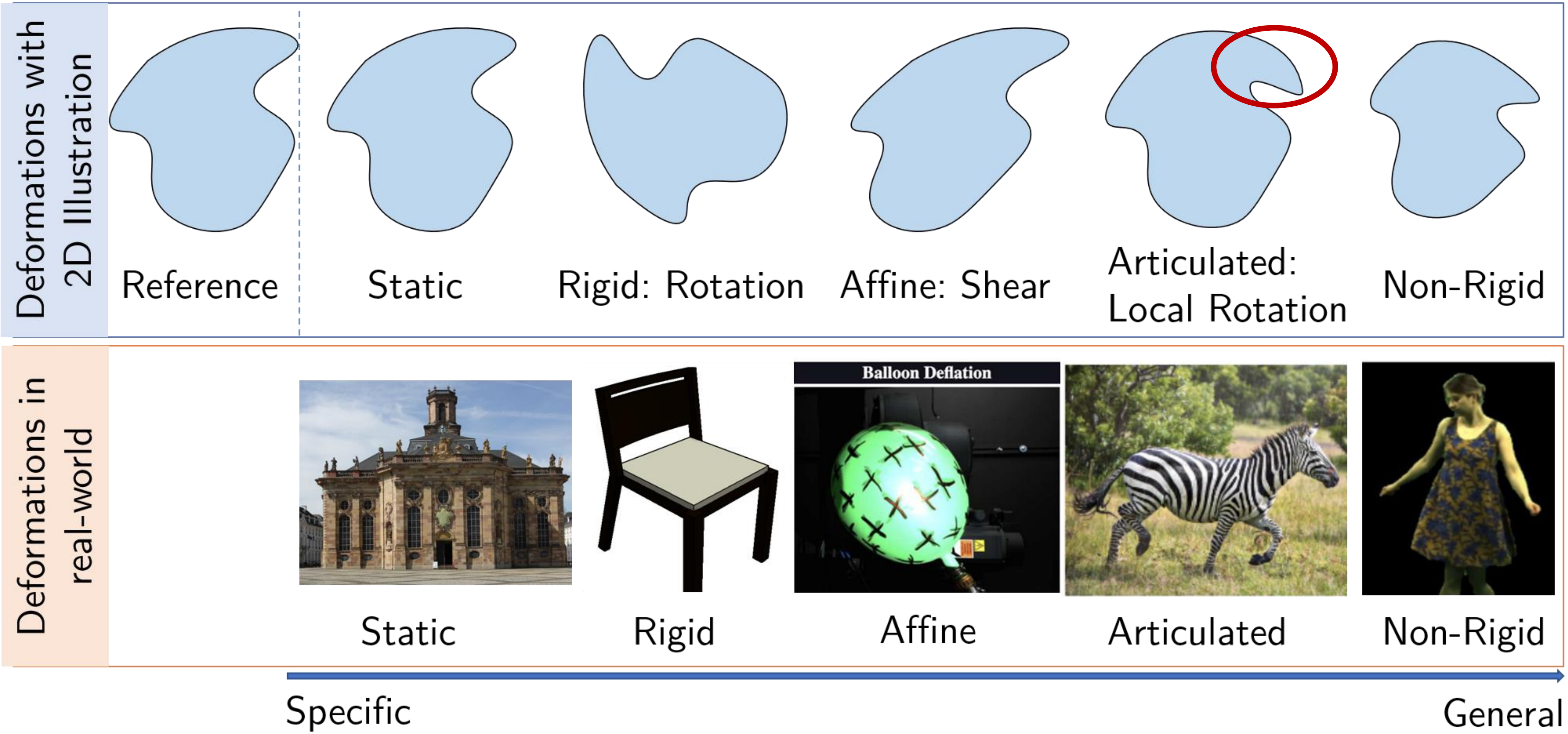
Appearance Functions



- Usually, simple models:

Appearance	Changes with viewing direction?	Model
Diffuse	No	Albedo: $\boldsymbol{c}(\mathbf{x})$
Glossy/specular	Yes	Simplified: $\boldsymbol{c}(\mathbf{x}, \mathbf{d})$ Full (BRDF): $c(\mathbf{x}, \mathbf{d}, \mathbf{l})$

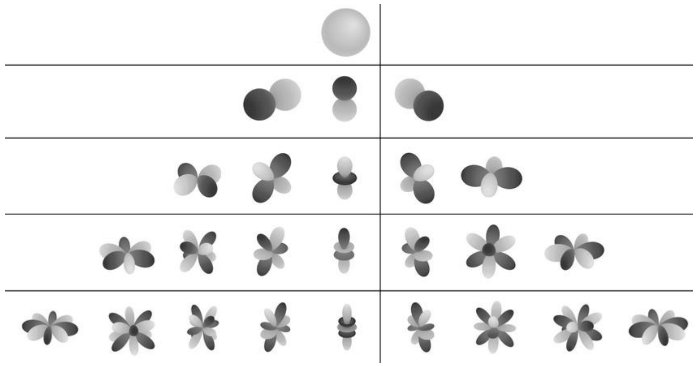
Deformation Categories



Geometry and Appearance Parametrizations

- Geometry:
  - Classically:
    - Point clouds and meshes as samples of the indicator function $s(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in S \\ 0 & \text{else} \end{cases}$
    - Voxel grids for level sets and densities
  - Neural:
    - Multi-layer perceptrons (MLPs) for levels sets and densities, *e.g.* $\text{density}(\mathbf{x}) = \text{MLP}(\mathbf{x})$

- Appearance:
  - Attach to each local unit, *e.g.* vertex
  - If appearance is not important, Lambertian model is used, *e.g.* RGB color per vertex
  - View dependence:
    - Classical: Spherical harmonics
    - Neural: $\boldsymbol{c}(\mathbf{x}, \mathbf{d}) = \text{MLP}(\mathbf{x}, \mathbf{d})$

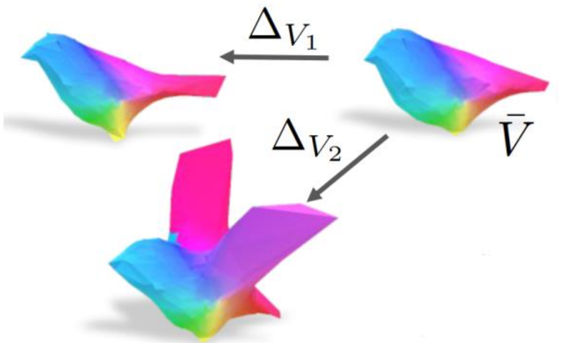


Deformation Parametrizations

- Ideally: Physics simulation
  - But: Difficult to model completely and computationally expensive

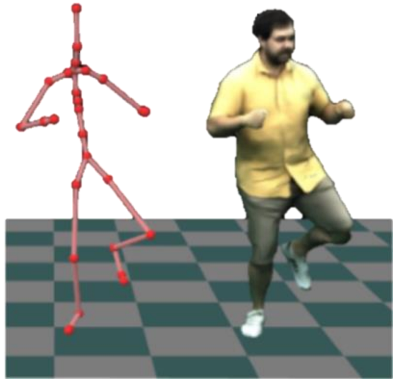
  → Non-physical approximations:

Template Offsets



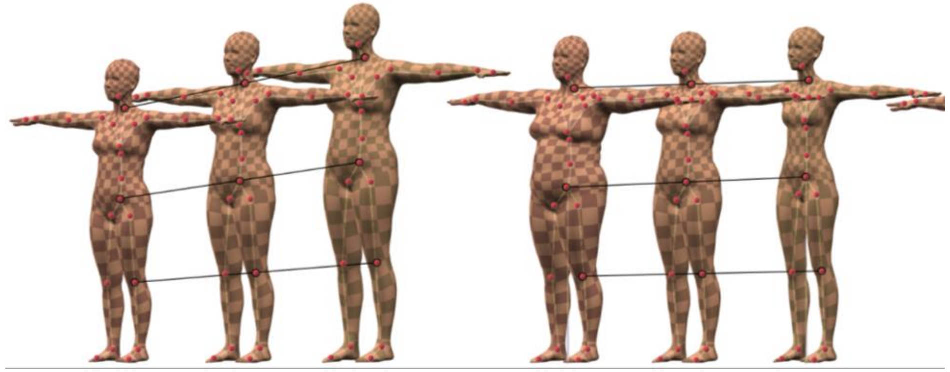
Kanazawa *et al.* 2018

Skinning



Habermann *et al.* 2021

Linear Subspace Models/
Parametric Models/3DMMs



Loper *et al.* 2015

Overview

- Representing deformations
- Rendering and data terms
- Challenges and priors to tackle them

Dense Monocular Non-Rigid Reconstruction

- Inverse and ill-posed problem
- Data term: Infinitely many solutions!
- Additional prior: Constrain the solution space

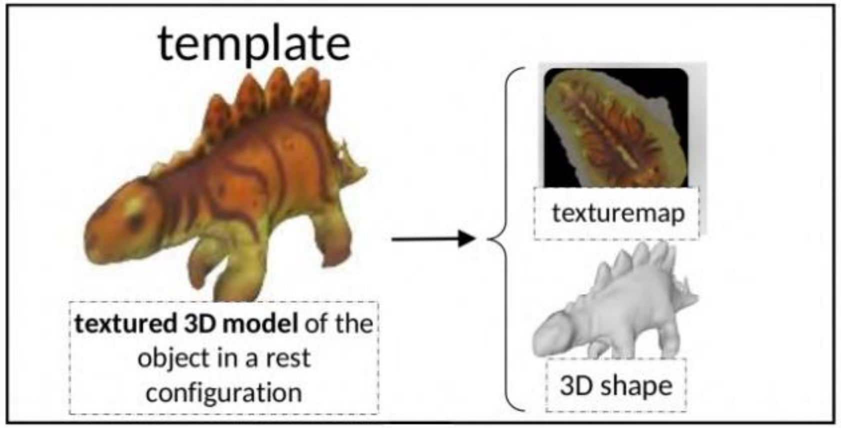
General method structure?

$$\mathcal{L}(\theta) = \boldsymbol{\mathcal{L}_{data}}(\boldsymbol{\theta}) + \mathcal{L}_{prior}(\theta)$$

Habermann *et al.* 2020

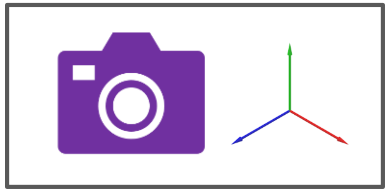
Saito *et al.* 2020

Data Terms: Additional Inputs

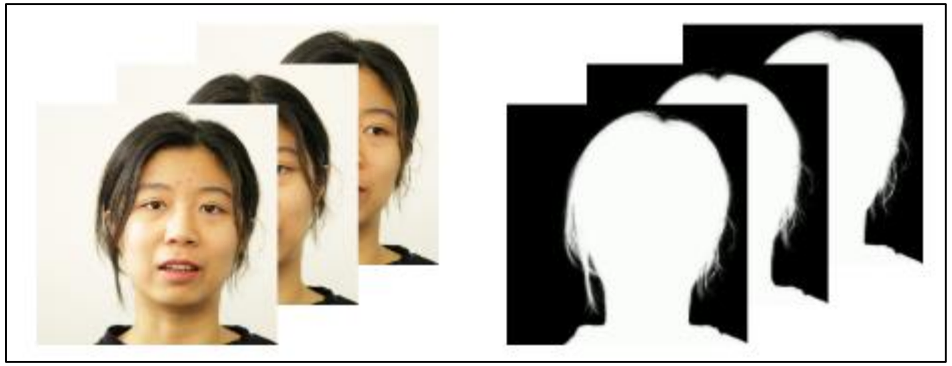


Fuentes-Jimenez *et al.* 2022

Camera

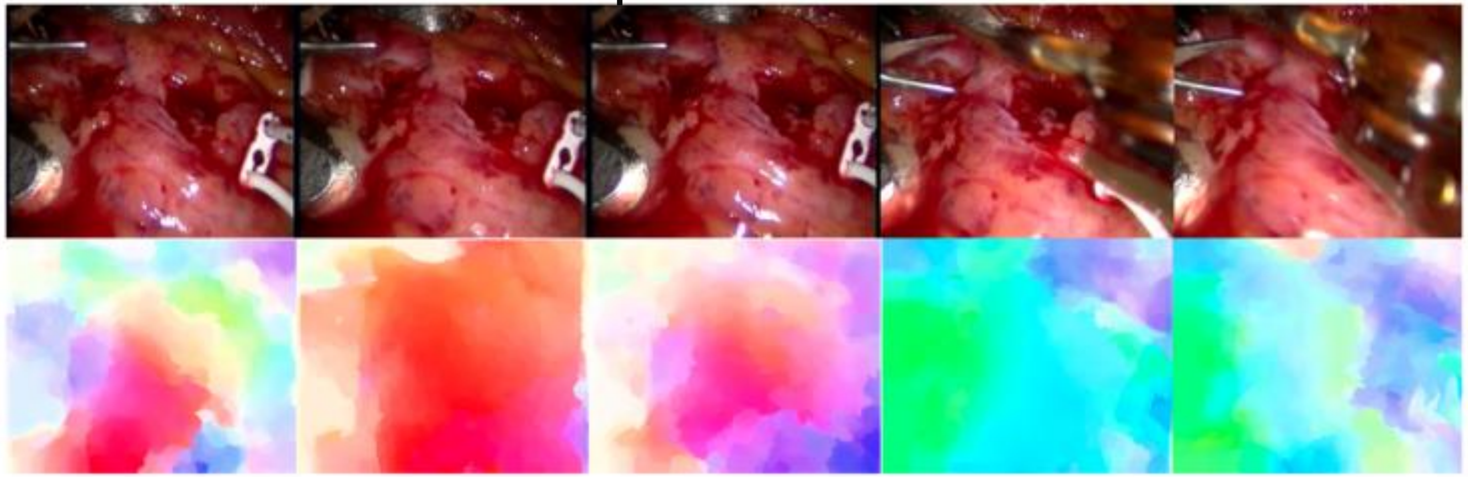


Segmentations



Zheng *et al.* 2022

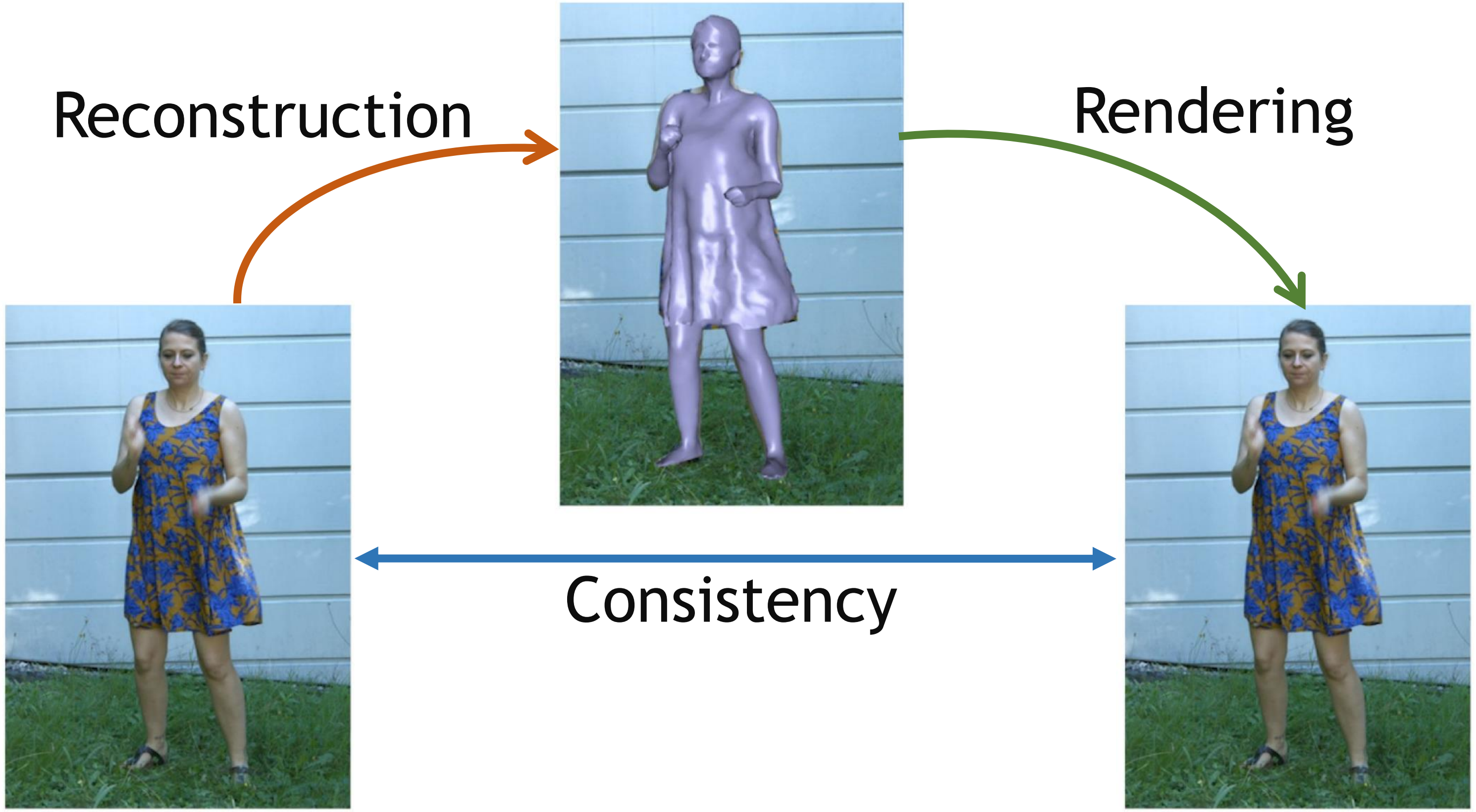
Optical Flow

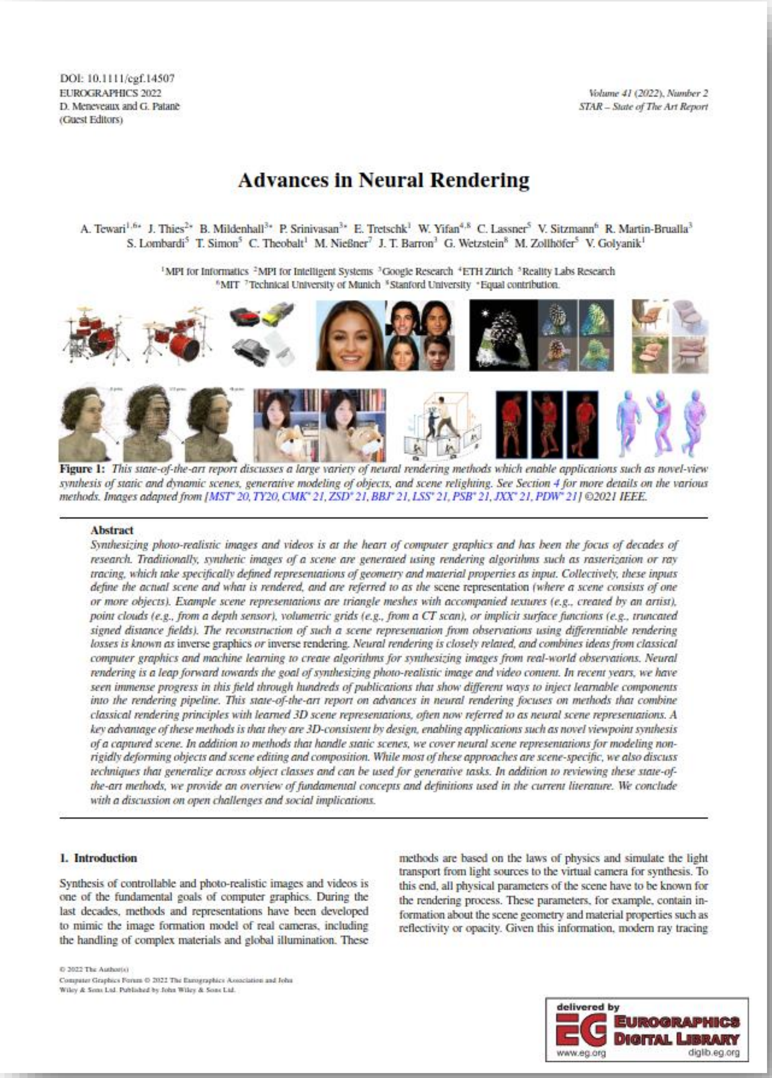


Golyanik *et al.* 2020

2D Keypoints

Data Terms: Rendering for 2D-3D Consistency




DOI: 10.1111/cgf.14507
EUROGRAPHICS 2022
D. Meneveaux and G. Patanè
(Guest Editors)

Volume 41 (2022), Number 2
STAR – State of The Art Report

Advances in Neural Rendering

A. Tewari[1,6*] J. Thies[2*] B. Mildenhall[3*] P. Srinivasan[3*] E. Tretschk[1] W. Yifan[4,8] C. Lassner[5] V. Sitzmann[6] R. Martin-Brualla[3] S. Lombardi[5] T. Simon[5] C. Theobalt[1] M. Nießner[7] J. T. Barron[3] G. Wetzstein[8] M. Zollhöfer[5] V. Golyanik[1]

[1]MPI for Informatics [2]MPI for Intelligent Systems [3]Google Research [4]ETH Zürich [5]Reality Labs Research [6]MIT [7]Technical University of Munich [8]Stanford University [*]Equal contribution.

Figure 1: *This state-of-the-art report discusses a large variety of neural rendering methods which enable applications such as novel-view synthesis of static and dynamic scenes, generative modeling of objects, and scene relighting. See Section 4 for more details on the various methods. Images adapted from [MST*20, TY20, CMK*21, ZSD*21, BBJ*21, LSS*21, PSB*21, JXX*21, PDW*21] ©2021 IEEE.*

Abstract

Synthesizing photo-realistic images and videos is at the heart of computer graphics and has been the focus of decades of research. Traditionally, synthetic images of a scene are generated using rendering algorithms such as rasterization or ray tracing, which take specifically defined representations of geometry and material properties as input. Collectively, these inputs define the actual scene and what is rendered, and are referred to as the scene representation *(where a scene consists of one or more objects). Example scene representations are triangle meshes with accompanied textures (e.g., created by an artist), point clouds (e.g., from a depth sensor), volumetric grids (e.g., from a CT scan), or implicit surface functions (e.g., truncated signed distance fields). The reconstruction of such a scene representation from observations using differentiable rendering losses is known as* inverse graphics *or* inverse rendering. *Neural rendering is closely related, and combines ideas from classical computer graphics and machine learning to create algorithms for synthesizing images from real-world observations. Neural rendering is a leap forward towards the goal of synthesizing photo-realistic image and video content. In recent years, we have seen immense progress in this field through hundreds of publications that show different ways to inject learnable components into the rendering pipeline. This state-of-the-art report on advances in neural rendering focuses on methods that combine classical rendering principles with learned 3D scene representations, often now referred to as neural scene representations. A key advantage of these methods is that they are 3D-consistent by design, enabling applications such as novel viewpoint synthesis of a captured scene. In addition to methods that handle static scenes, we cover neural scene representations for modeling non-rigidly deforming objects and scene editing and composition. While most of these approaches are scene-specific, we also discuss techniques that generalize across object classes and can be used for generative tasks. In addition to reviewing these state-of-the-art methods, we provide an overview of fundamental concepts and definitions used in the current literature. We conclude with a discussion on open challenges and social implications.*

1. Introduction

Synthesis of controllable and photo-realistic images and videos is one of the fundamental goals of computer graphics. During the last decades, methods and representations have been developed to mimic the image formation model of real cameras, including the handling of complex materials and global illumination. These methods are based on the laws of physics and simulate the light transport from light sources to the virtual camera for synthesis. To this end, all physical parameters of the scene have to be known for the rendering process. These parameters, for example, contain information about the scene geometry and material properties such as reflectivity or opacity. Given this information, modern ray tracing

© 2022 The Author(s)
Computer Graphics Forum © 2022 The Eurographics Association and John Wiley & Sons Ltd. Published by John Wiley & Sons Ltd.

delivered by EUROGRAPHICS DIGITAL LIBRARY www.eg.org diglib.eg.org

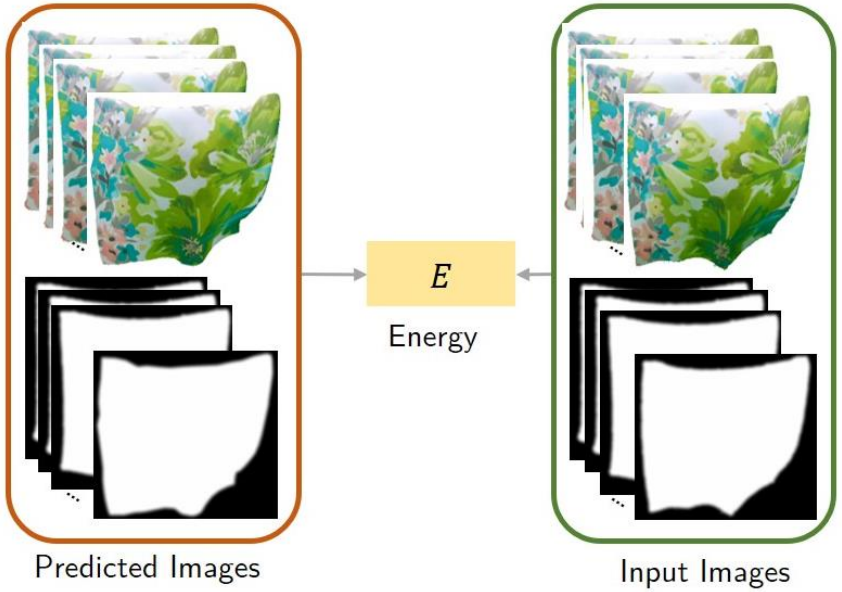
Tewari *et al.* 2022

Differentiable rendering?

→ Differentiable rasterization, volume rendering

Data Terms: 2D-3D Consistency

Photometric Consistency

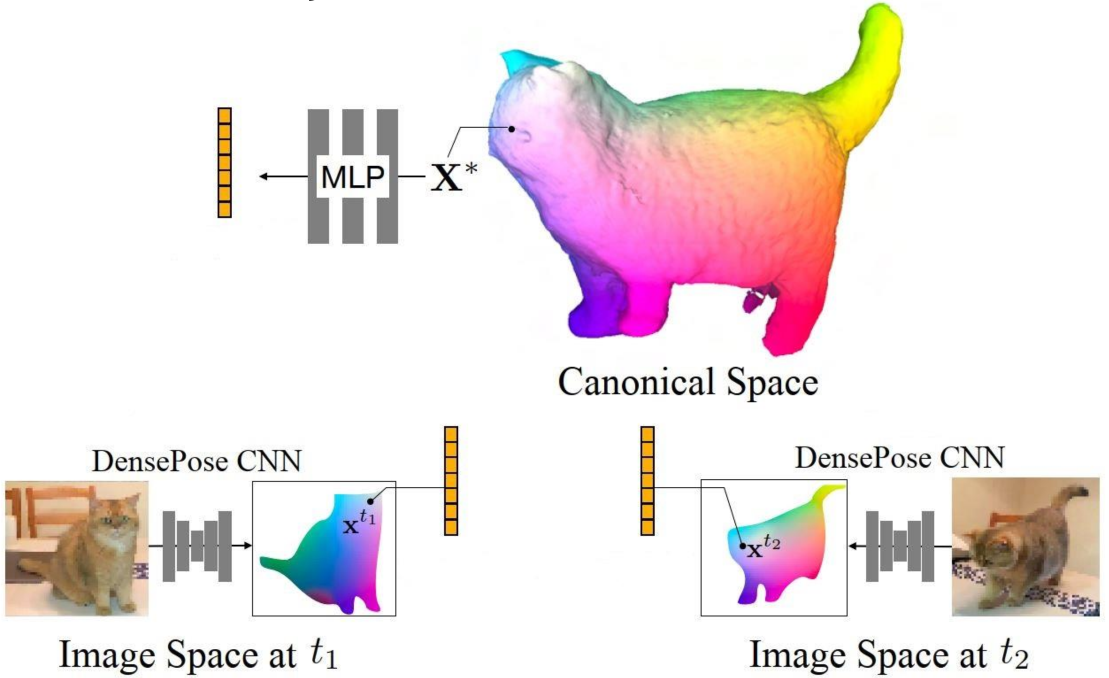


Perceptual Consistency

	Patch 0	Reference	Patch 1
L2			✓
Human	✓		
LPIPS	✓		

Zhang *et al.* 2018

Consistency of Learned Features



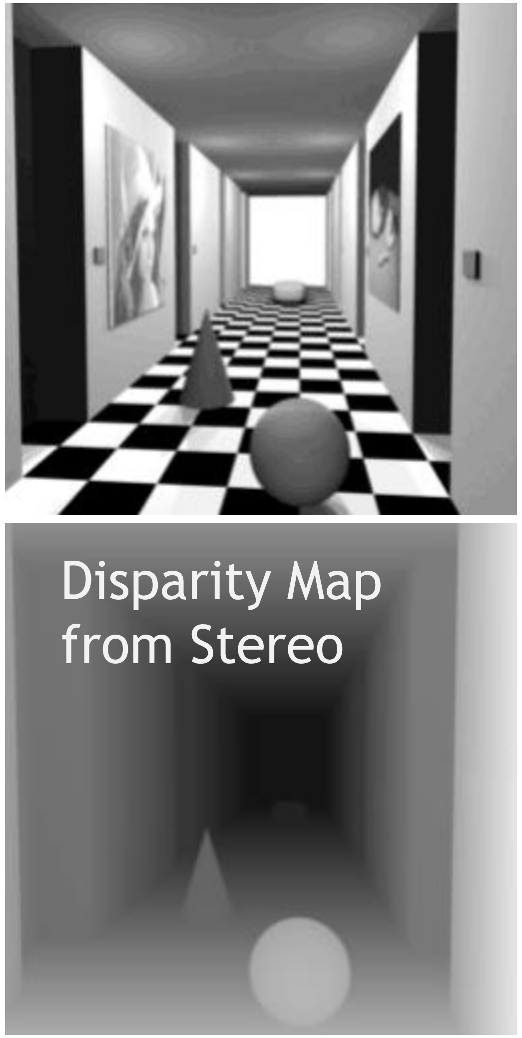
Yang *et al.* 2022

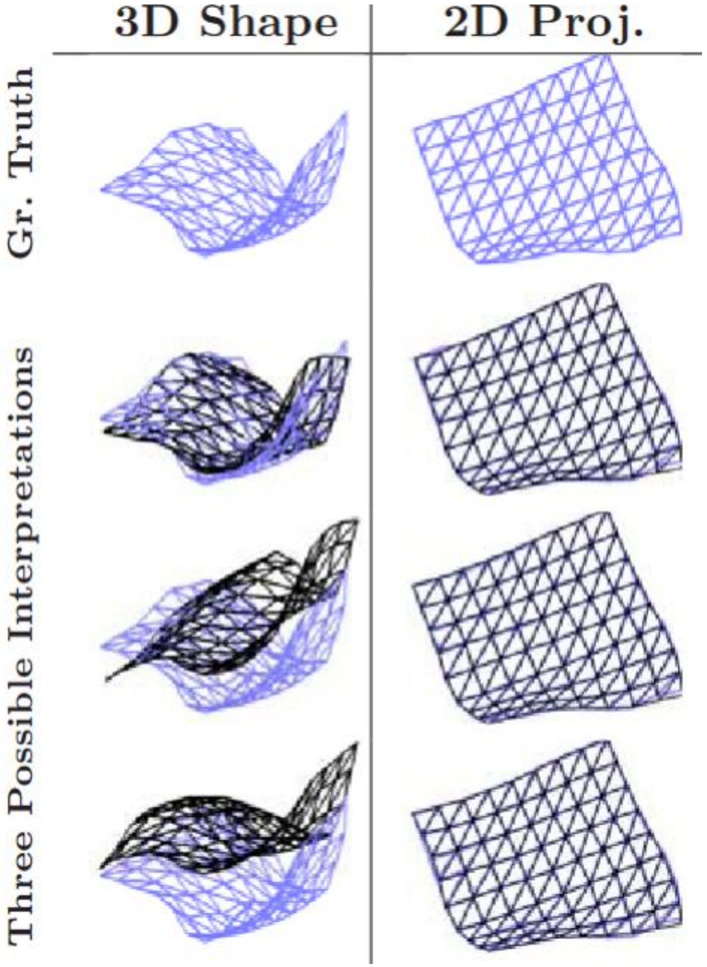
Overview

- Representing deformations
- Rendering and data terms
- Challenges and priors to tackle them

Reconstruction: Inherent Challenges

Monocular Depth Ambiguity





Moreno-Noguer *et al*. 2010

Attributing fine-scale details to geometry vs. appearance?

Chan *et al*. 2022

Reconstruction: Inherent Challenges

Occlusion

View-dependence

Verbin *et al.* 2022

Reconstruction: Parameterization Challenges

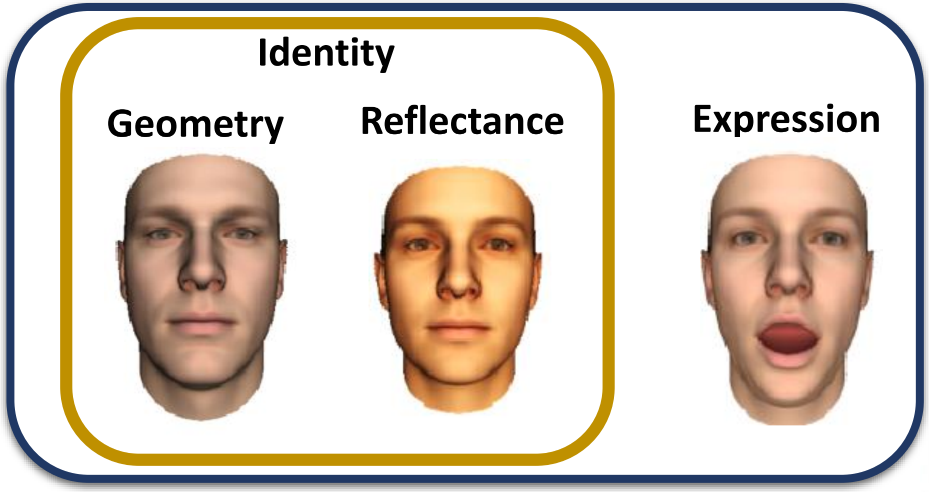
Topology Change

Explicit meshes (topology changes are challenging)
Li *et al*. 2021

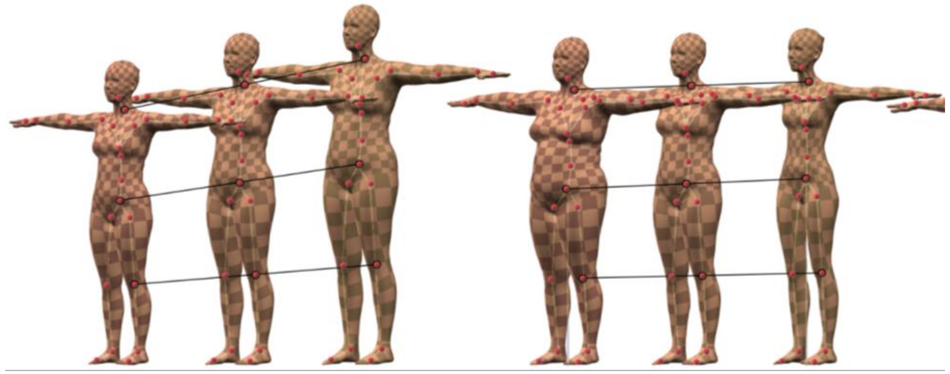
Implicit functions
(no correspondences)
Saito *et al*. 2020

Reconstruction: Parameterization Challenges

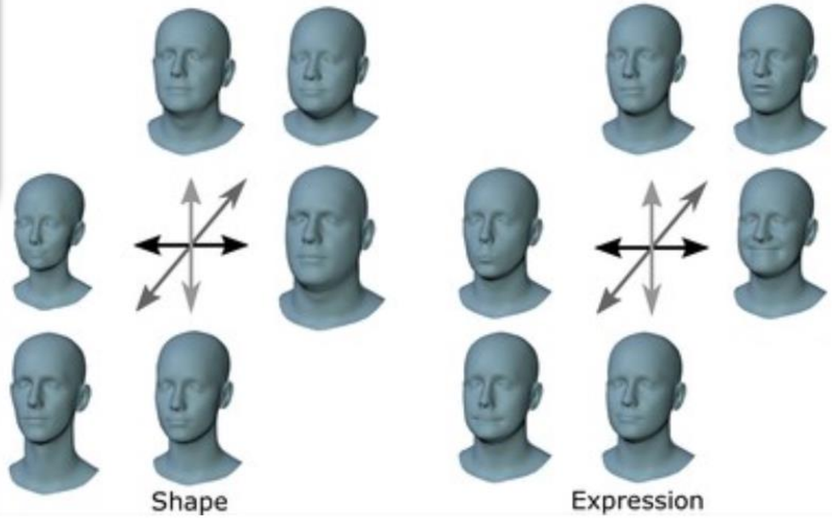
Identity-Deformation Ambiguity



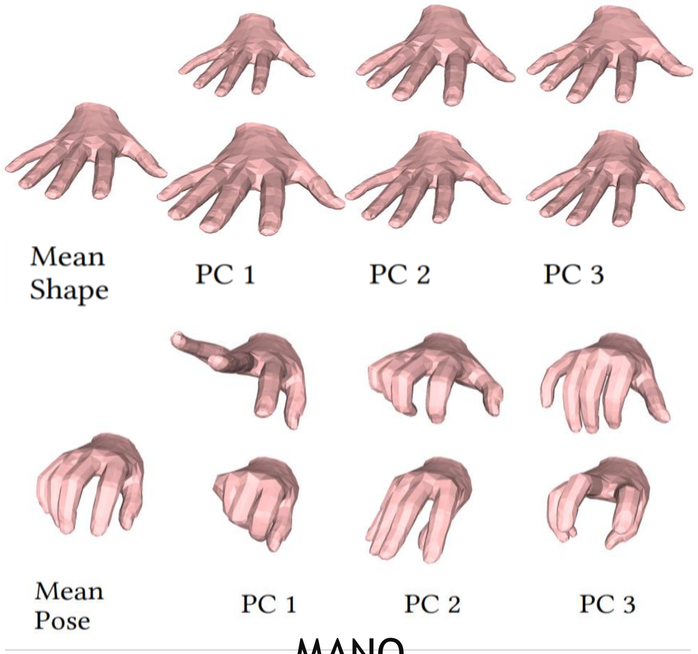
Tewari *et al*. 2017



SMPL



FLAME



MANO

Reconstruction: Data Acquisition Challenges

Background

Motion Blur

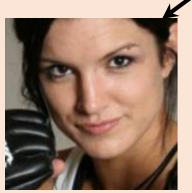
Rudnev *et al.* 2021

Dense Monocular Non-Rigid Reconstruction

- Ill-posed inverse problem
- Data term: Infinitely many solutions!
- Additional prior: Constrain the solution space

General method structure?

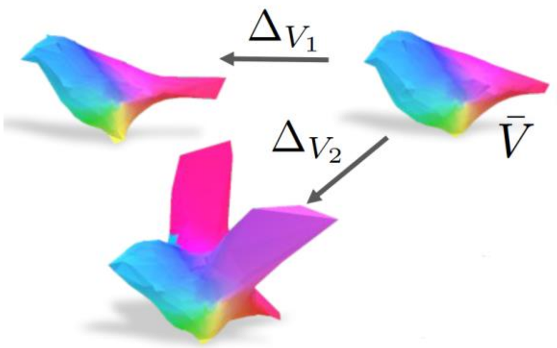
$$\mathcal{L}(\theta) = \mathcal{L}_{data}(\theta) + \boldsymbol{\mathcal{L}}_{\boldsymbol{prior}}(\boldsymbol{\theta})$$

$\Big\|$  $-$  $\Big\|_2$

Priors

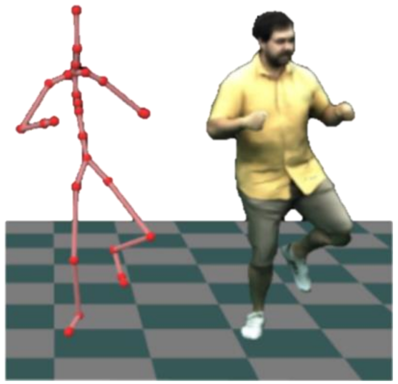
Hard priors: Geometry parameterization

Template Offsets



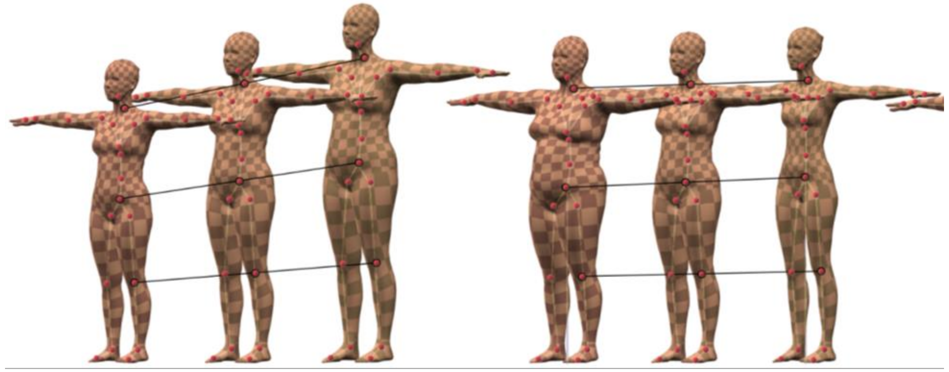
Kanazawa *et al.* 2018

Skinning



Habermann *et al.* 2021

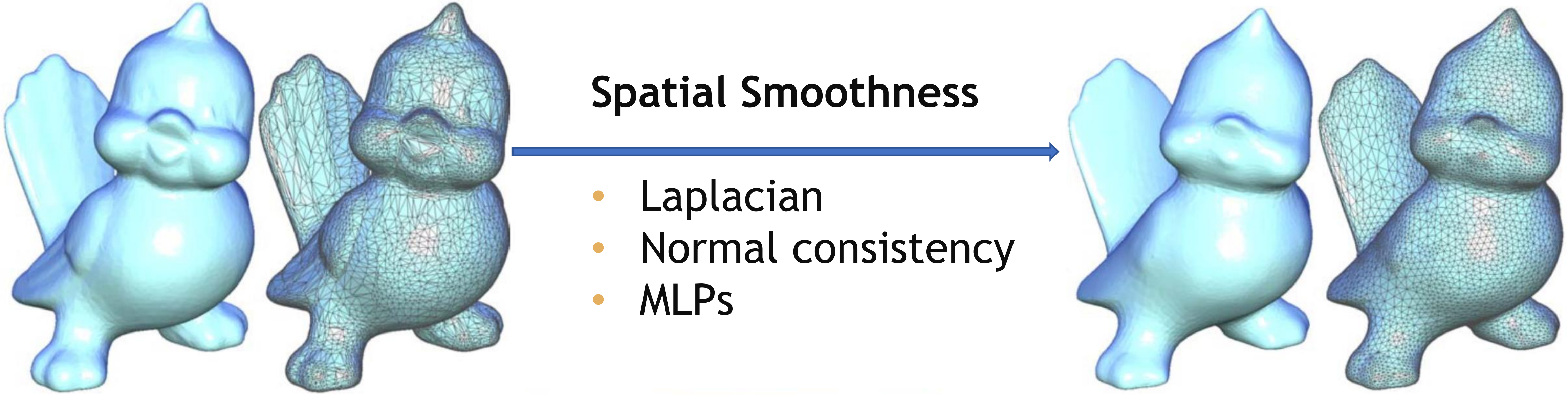
Linear Subspace Models/
Parametric Models/3DMMs



Loper *et al.* 2015

Soft priors? Next…

Geometry Soft Priors

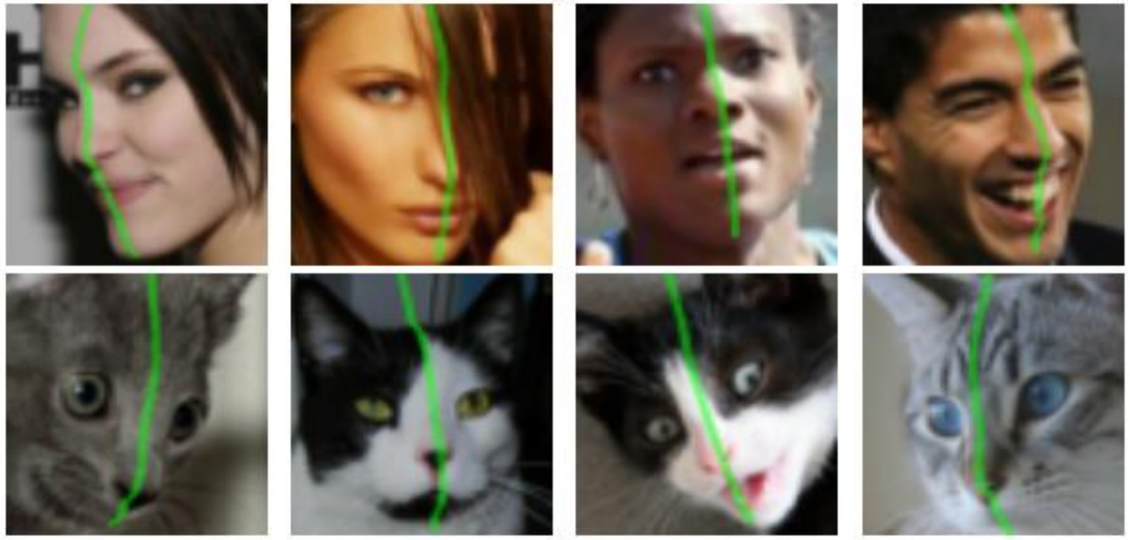


Spatial Smoothness

- Laplacian
- Normal consistency
- MLPs

Nealen *et al.* 2006

Symmetry Constraints

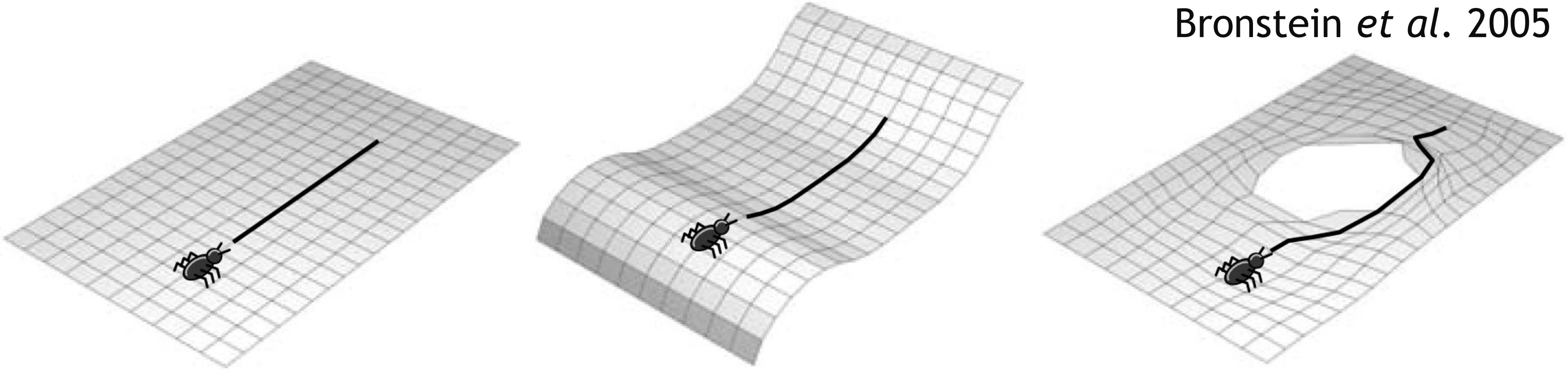


Wu *et al.* 2020

Deformation Soft Priors: Reference Geometry

Bronstein *et al.* 2005

Metric-based prior
- Isometric
- Conformal

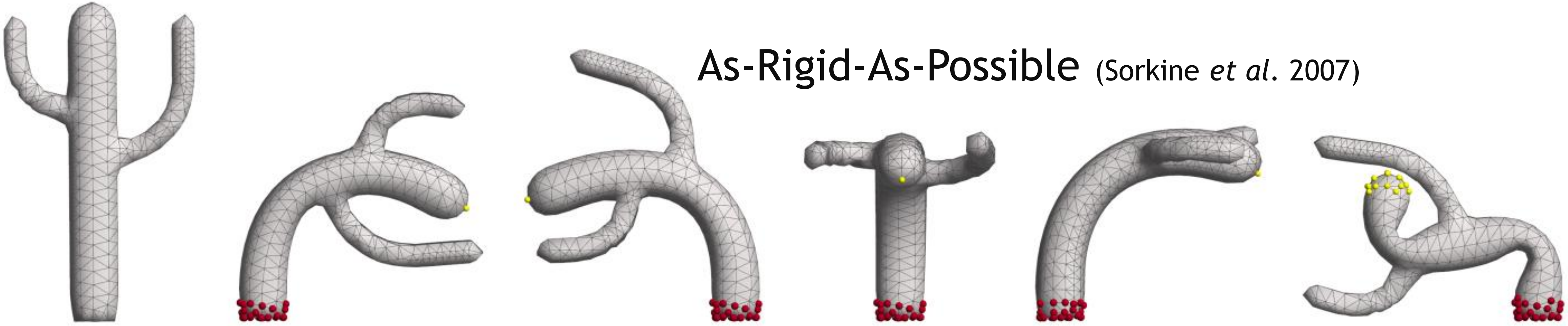


(a) Original surface

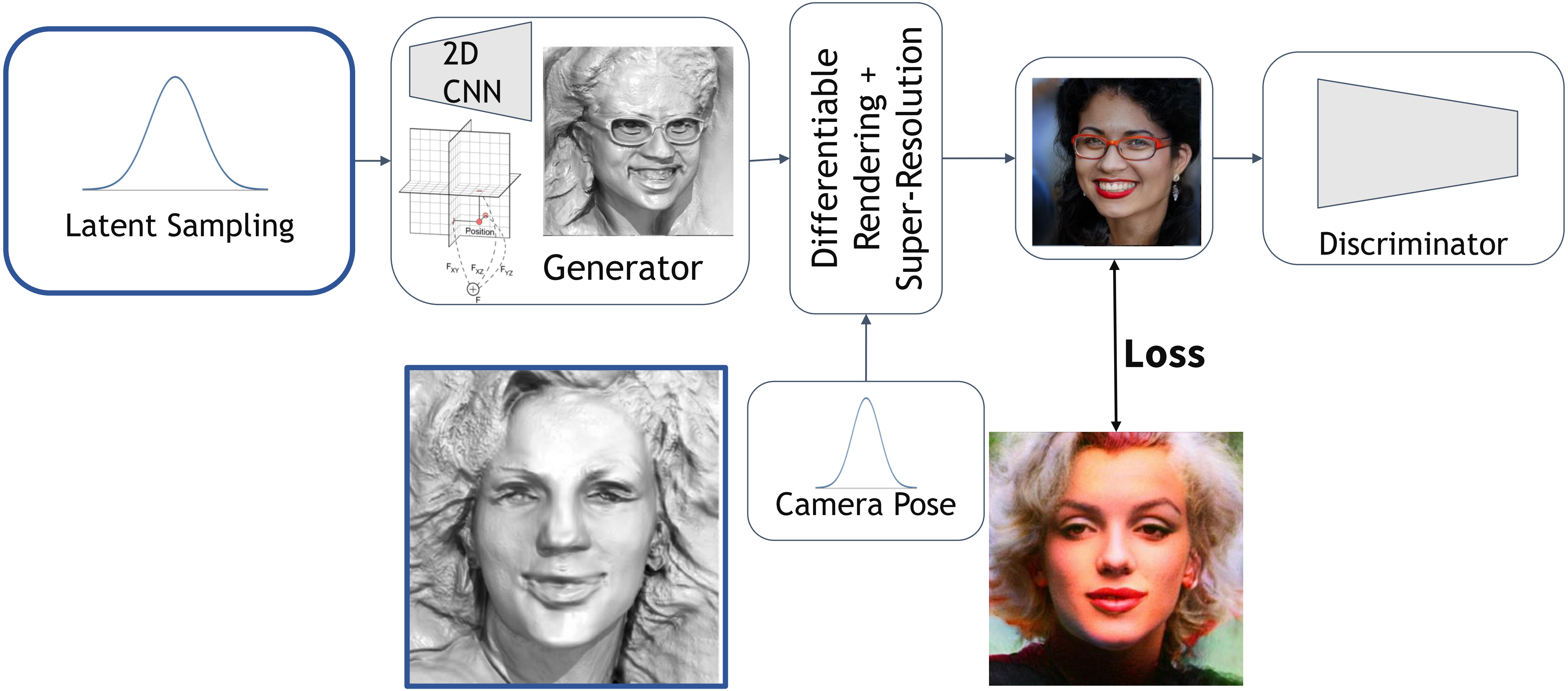
(b) Isometric transformation

(c) Non-isometric transformation

As-Rigid-As-Possible (Sorkine *et al.* 2007)



3D-Aware GAN Prior



Chan *et al.* 2022

Optimization: Finding the Right Parameters

Loss $\mathcal{L}(\theta) = \mathcal{L}_{data}(\theta) + \lambda\mathcal{L}_{prior}(\theta)$

Optimal parameters $\theta^* = \arg_\theta \min \mathcal{L}(\theta)$

Optimization: Gradient-based techniques

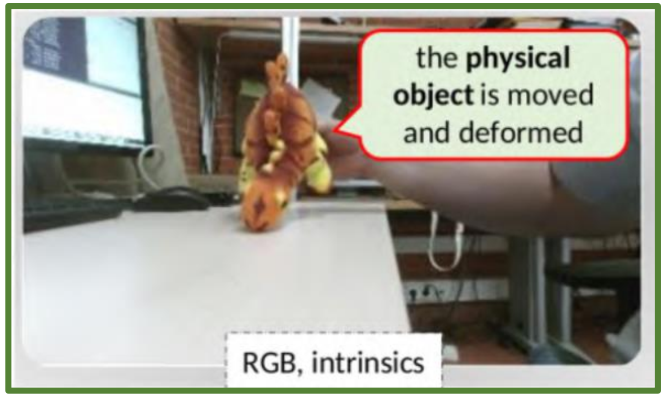
3
State-of-the-Art Methods

3.1
General Objects

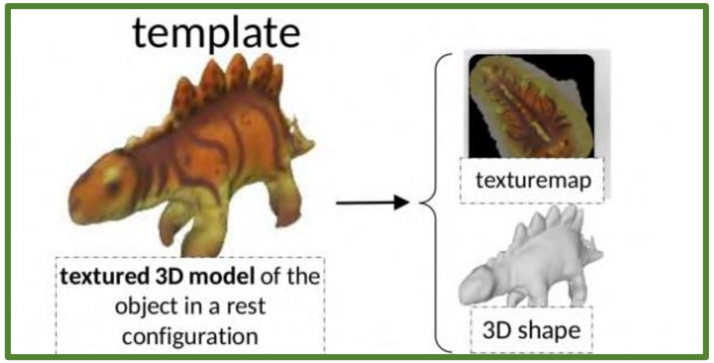
3.1.1
Shape from Template

Shape from Template

Input

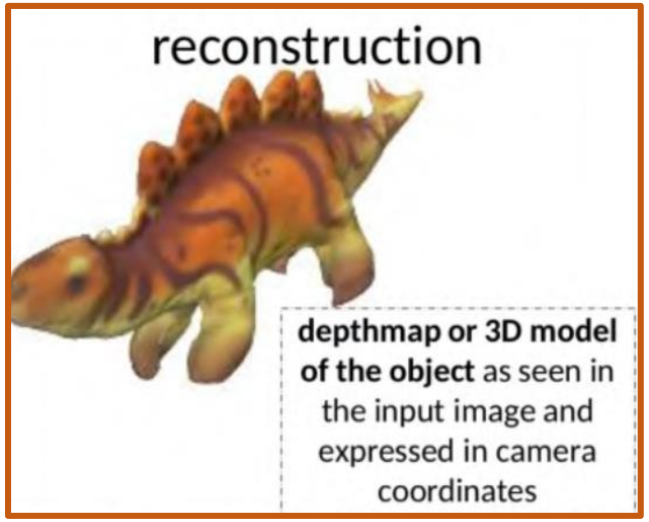


Single Image

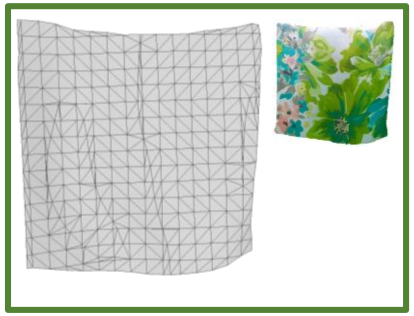


Template

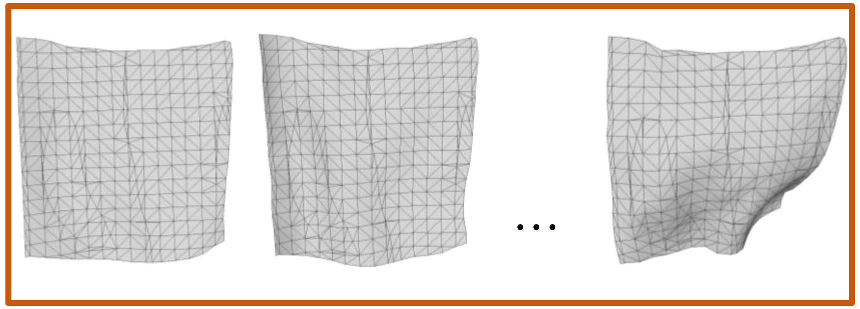
Output



Monocular Video

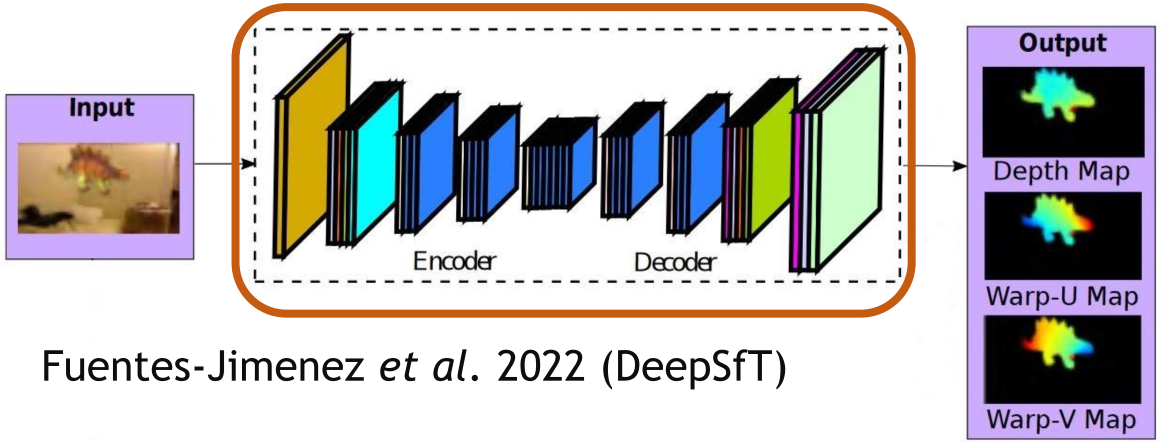


Template

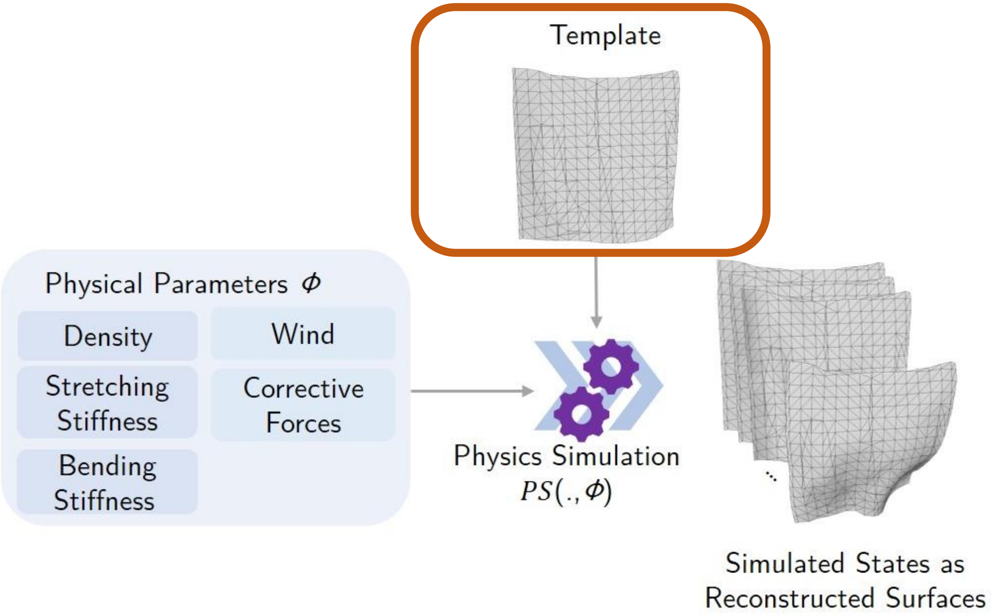


3D Reconstructions

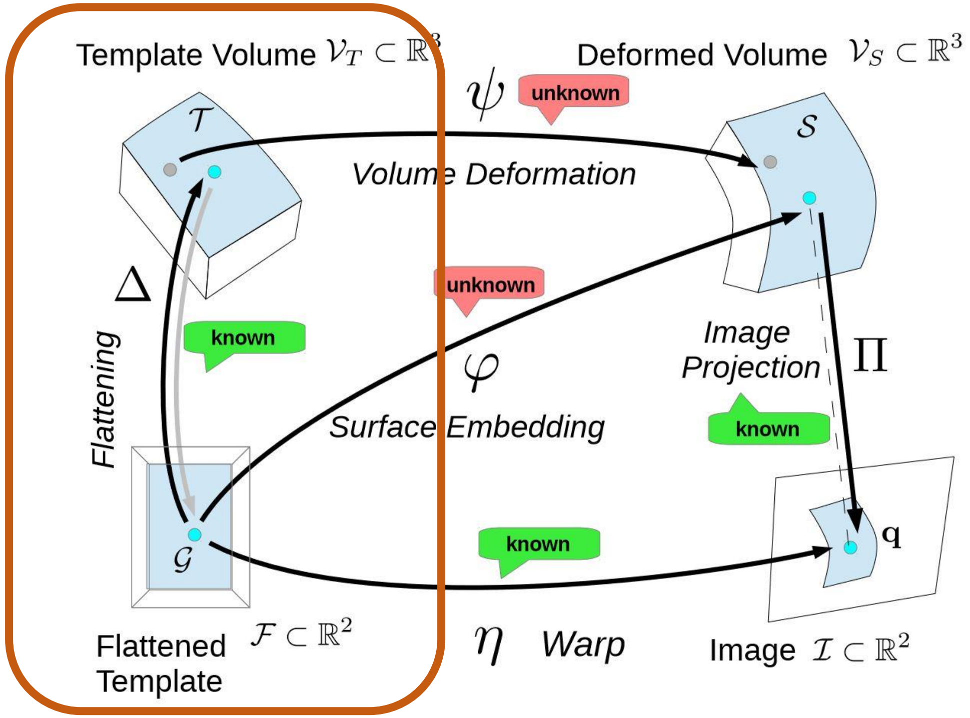
Shape from Template: How is the Template Used?



Fuentes-Jimenez *et al.* 2022 (DeepSfT)

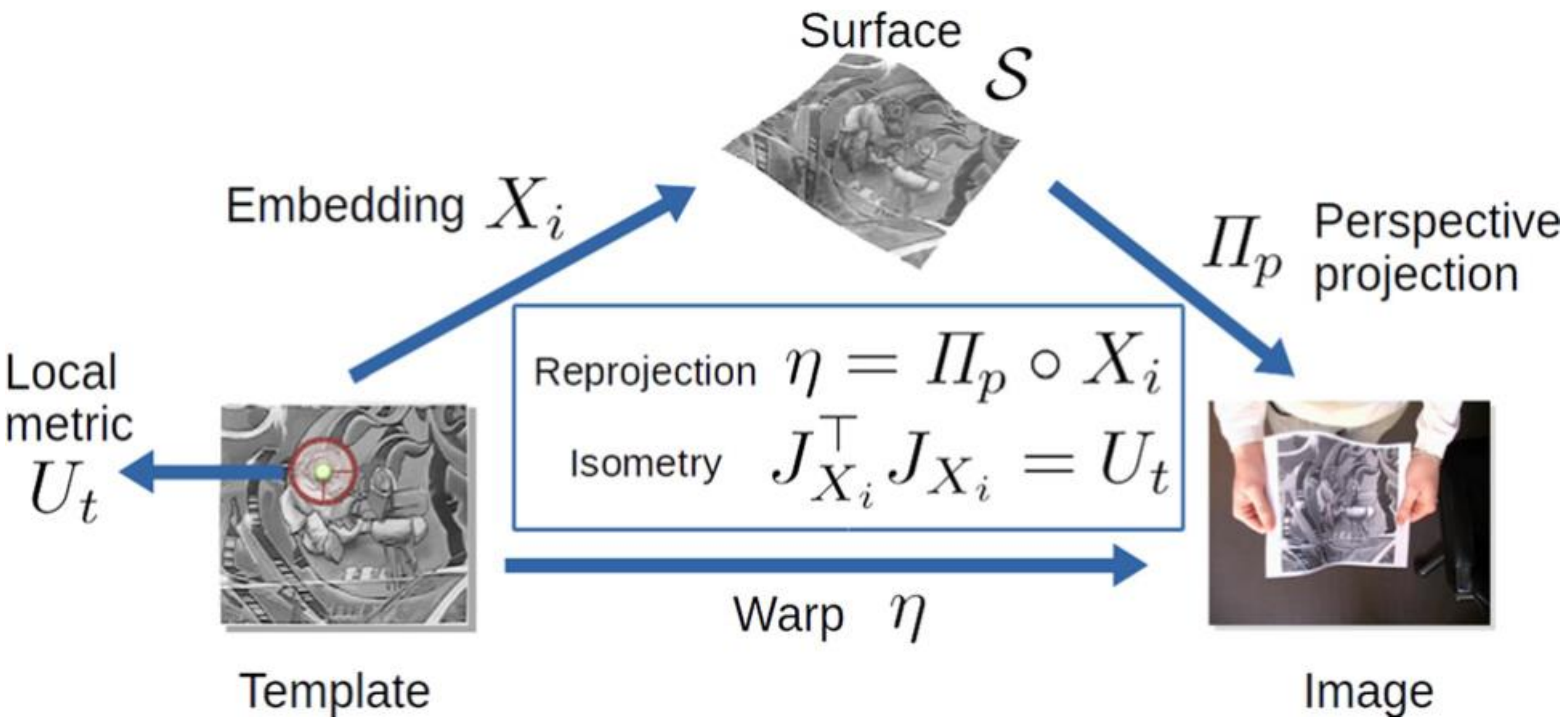


Kairanda *et al.* CVPR 2022 (φ-SfT)



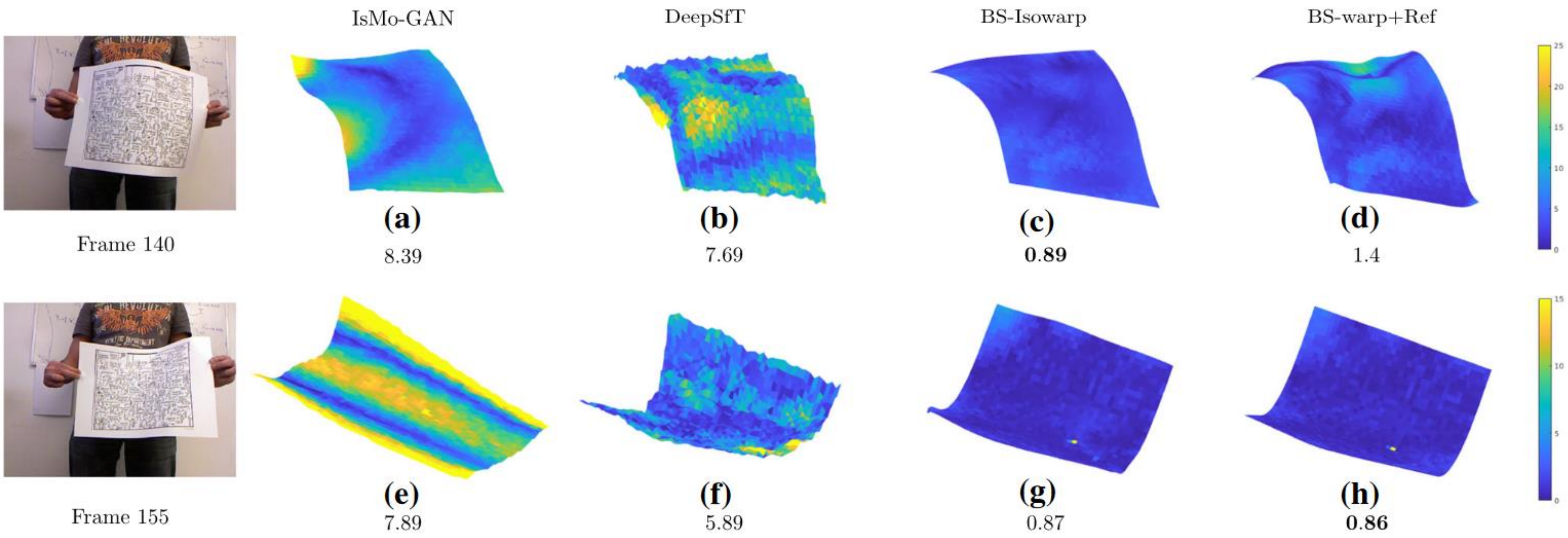
Parashar *et al.* 2015

State-of-the-Art SfT: Analytical Methods



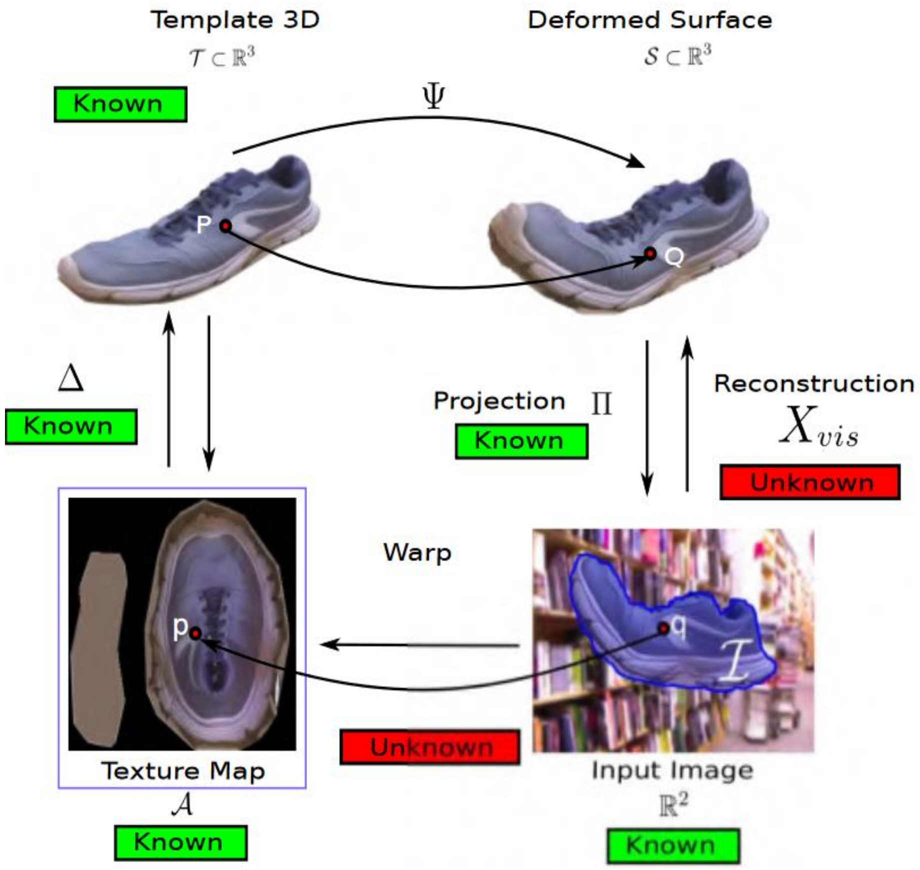
Casillas-Perez *et al*. 2021 (Isowarp)
Chhatkuli *et al*. 2016
Bartoli *et al*. 2015

State-of-the-Art SfT: Analytical Methods

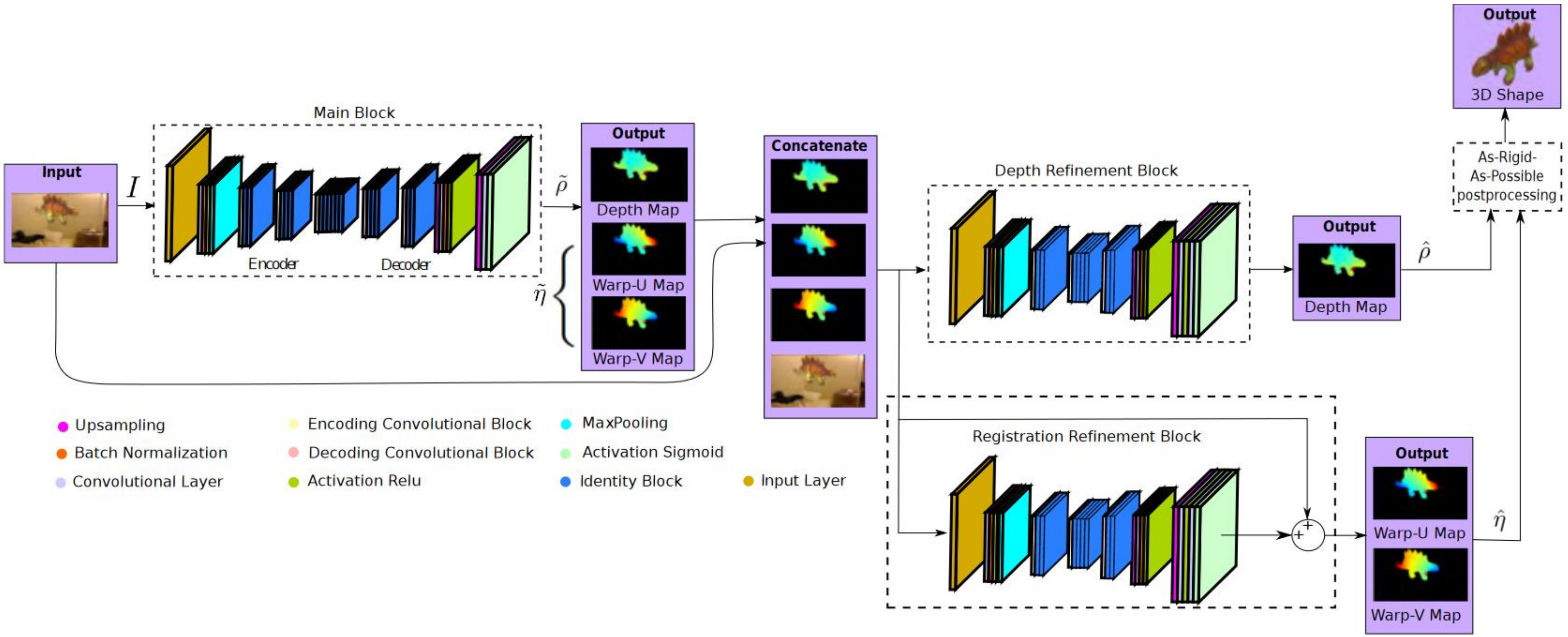


Casillas-Perez *et al*. 2021 (Isowarp)

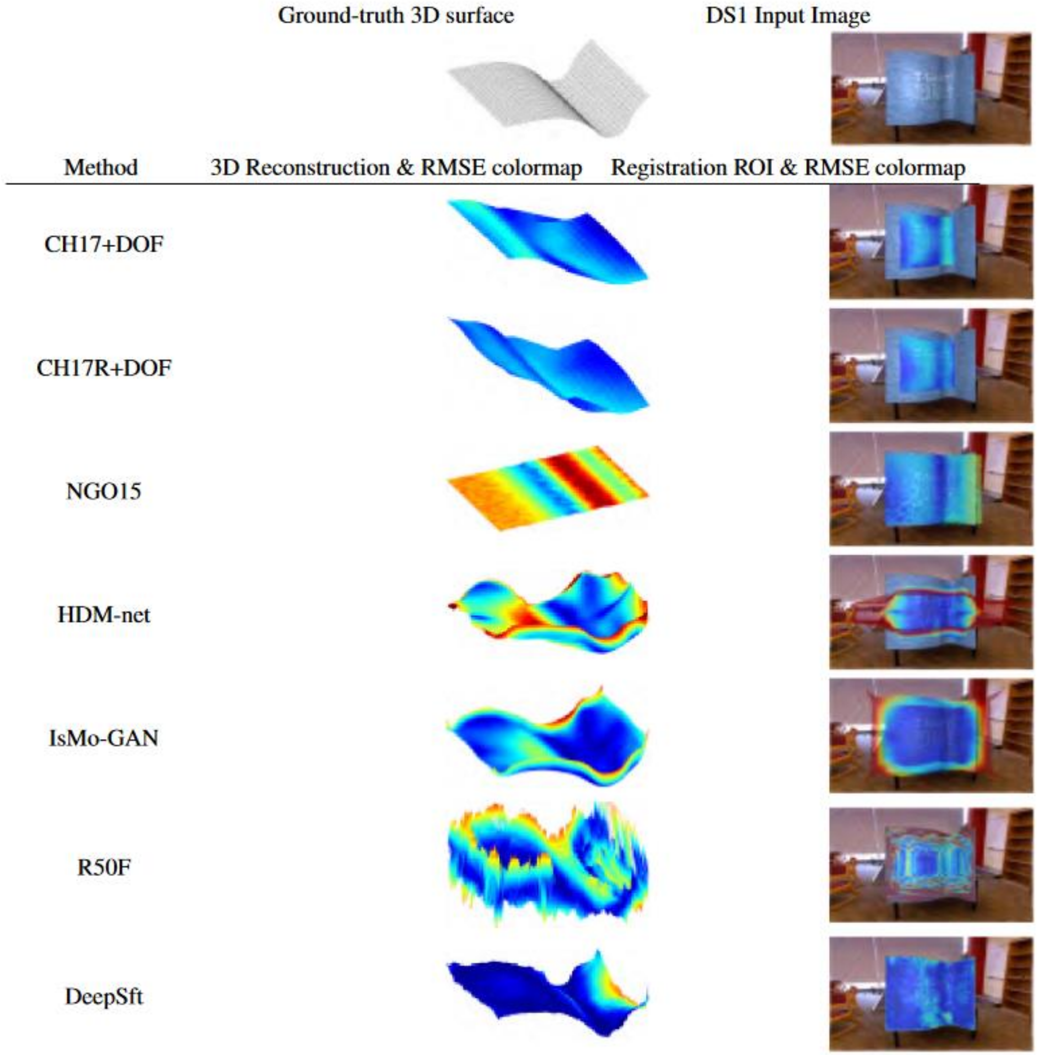
State-of-the-Art SfT: Neural Methods

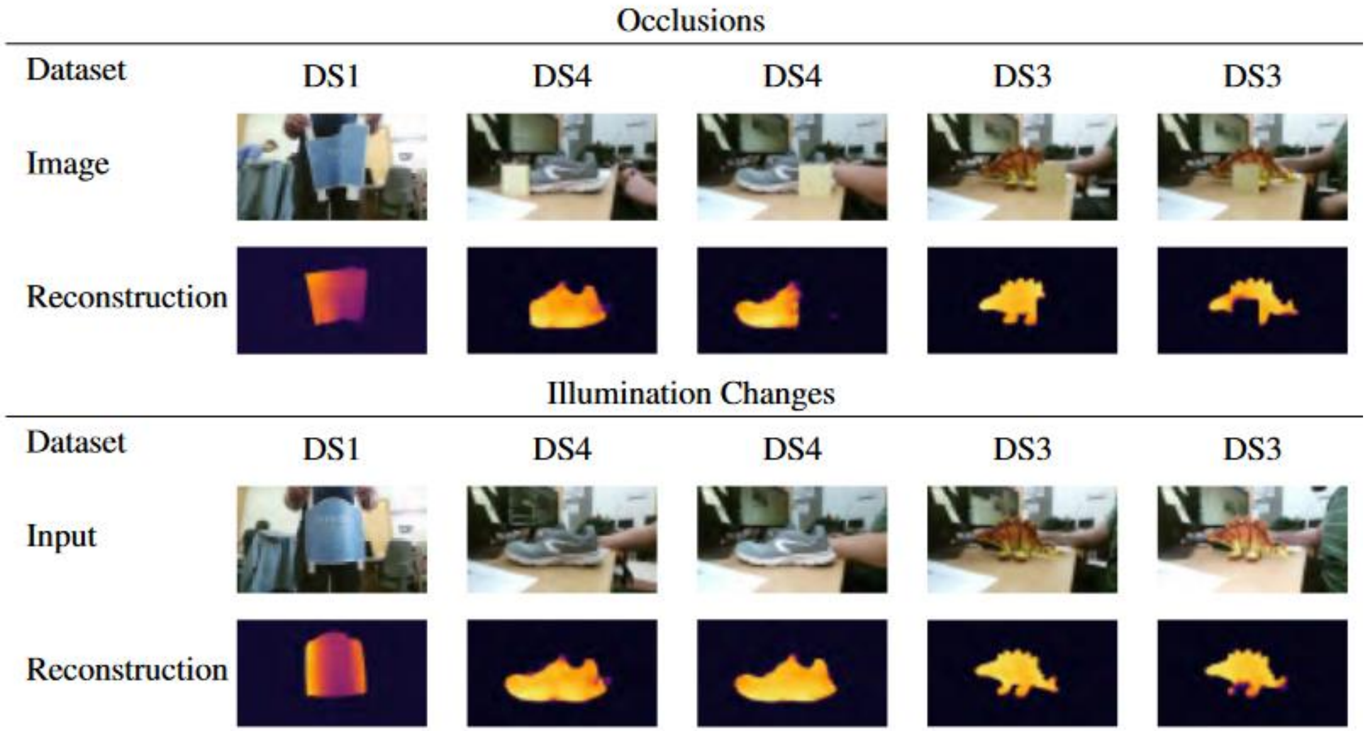


State-of-the-Art SfT: Neural Methods



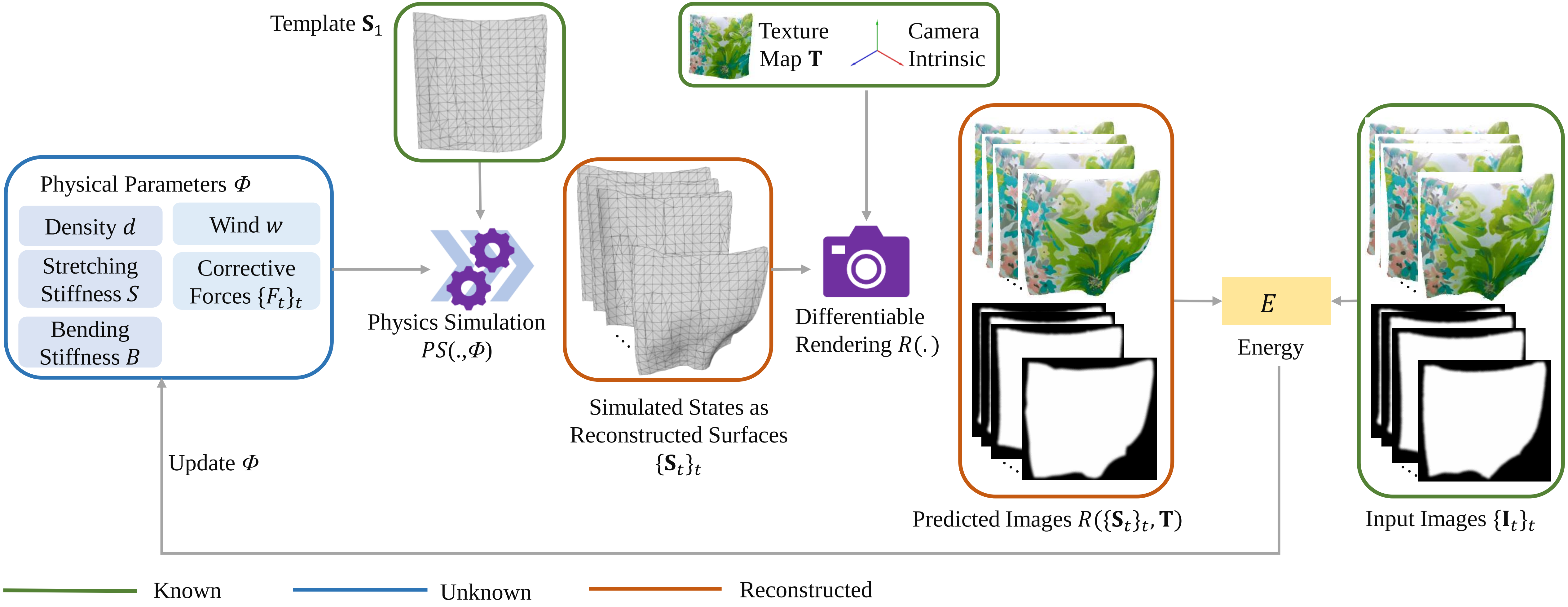
State-of-the-Art SfT: Neural Methods





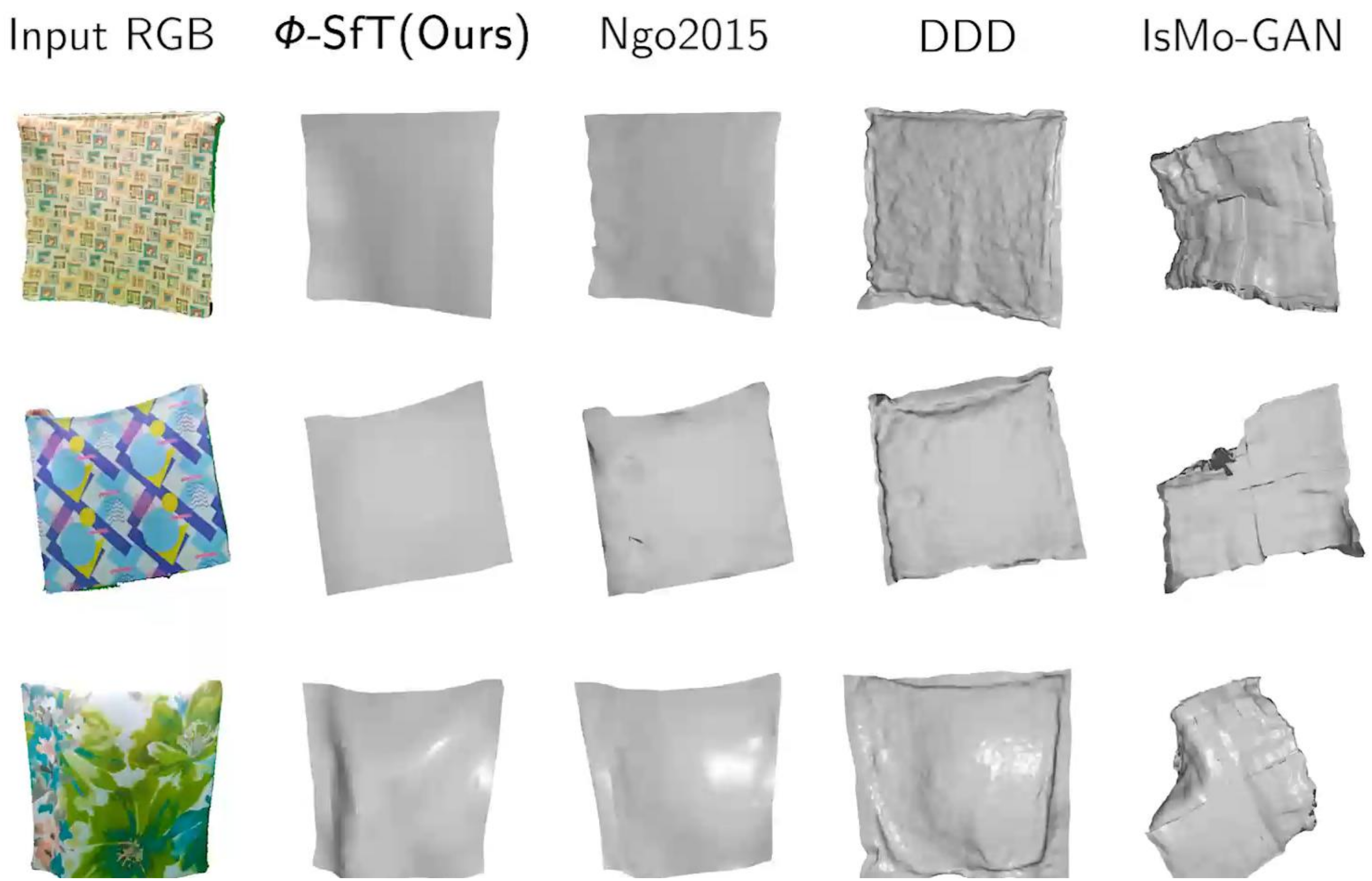
Fuentes-Jimenez *et al.* 2022 (DeepSfT)

State-of-the-Art SfT: Energy-Based Methods



Kairanda *et al*. 2022 (φ-SfT)

State-of-the-Art SfT: Energy-Based Methods



Kairanda *et al*. 2022 (φ-SfT)

Open Challenges

- Generalizability
  - Single deformable objects
  - Evaluated on smooth deformations
  - Missing background reconstruction
  - Changing object topology
  - Self-collision

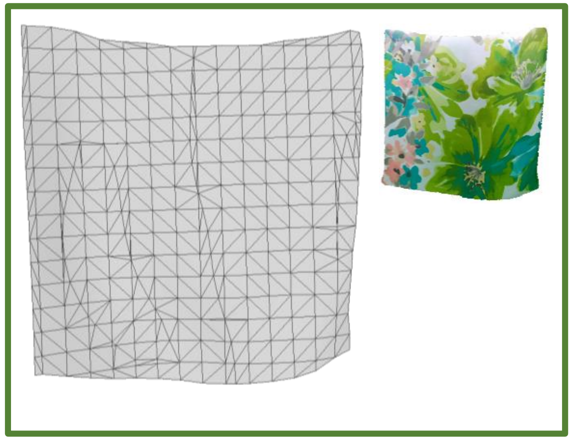
- Assumptions
  - Template availability
  - Errors in image-to-template warp

3.1.2 Non-Rigid Structure from Motion

Dense Non-Rigid Structure from Motion (NRSfM)

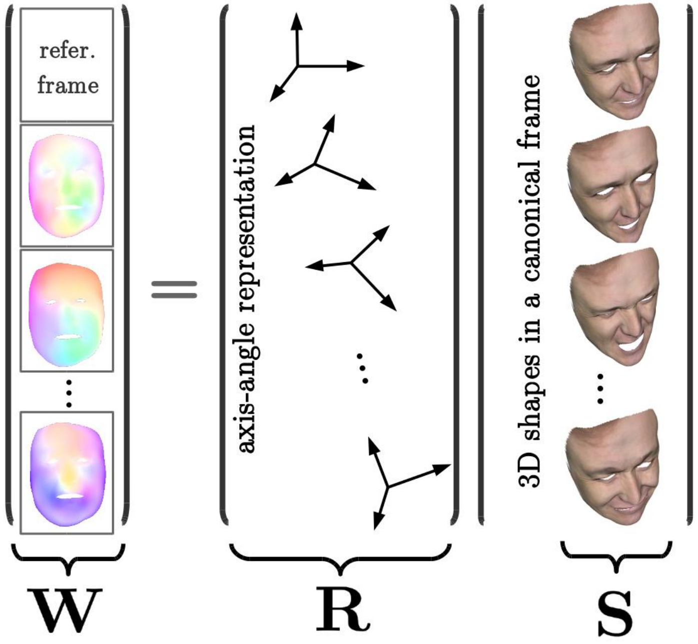
- Motion and deformation cues for 3D recovery
- Most SOTA methods follow the matrix factorization approach of Bregler *et al.*
- Prior assumption: Deformable shapes span low-rank subspaces

Shape from Template



Input: Image and 3D template

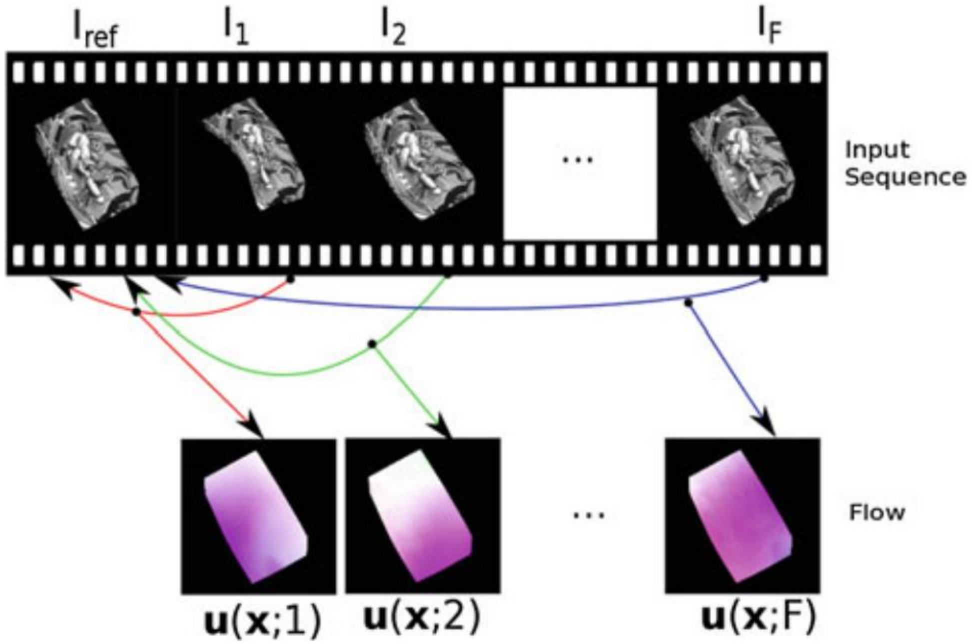
Non-Rigid Structure from Motion



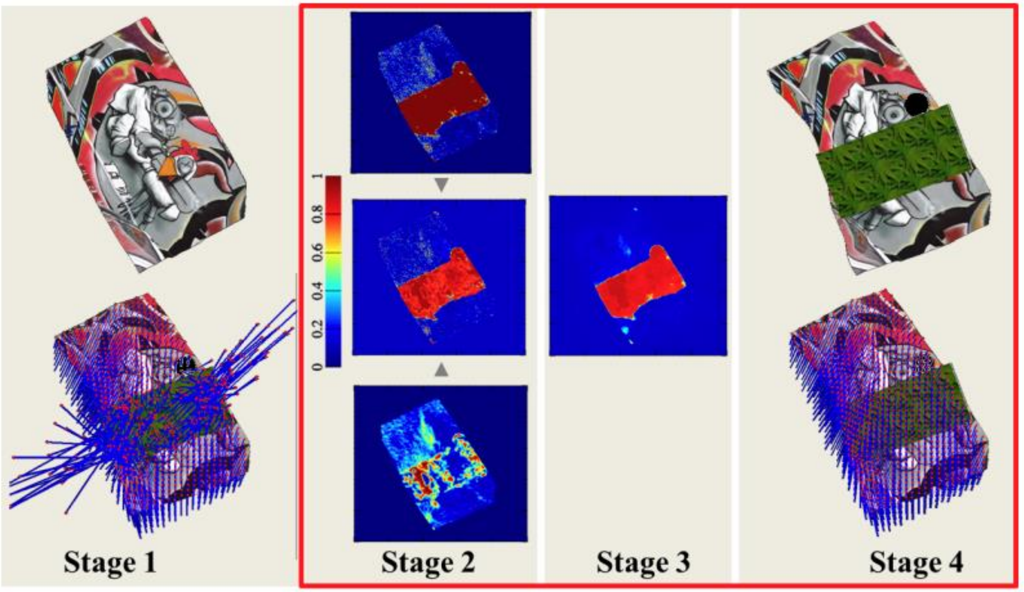
Input: 2D point tracks

Sidhu *et al.* 2020

Dense NRSfM: 2D Point Tracking



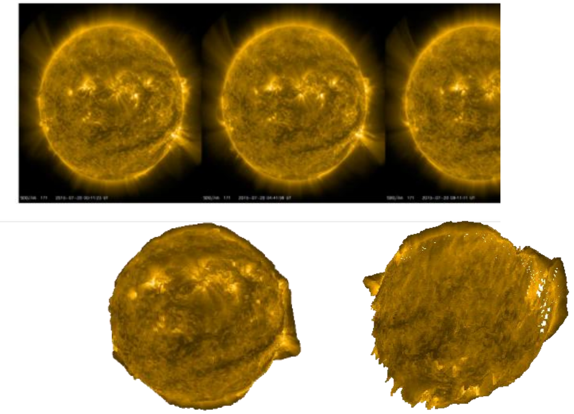
Garg *et al*. 2013:
Multi-frame optical flow / video registration



Taetz *et al*. 2016:
Occlusion-aware video registration

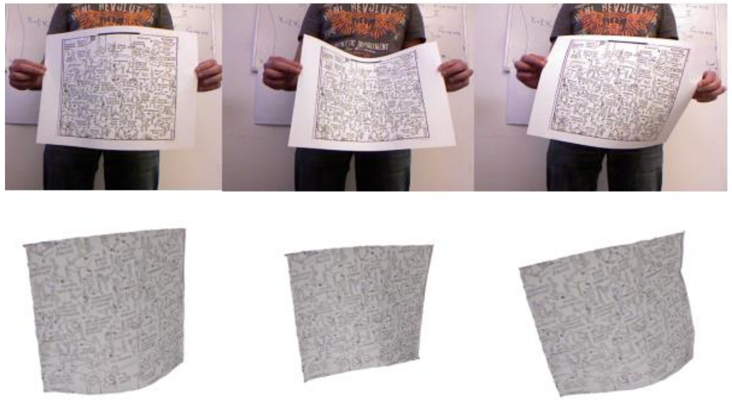
Occluded in the image

Dense NRSfM: Different Object Scales

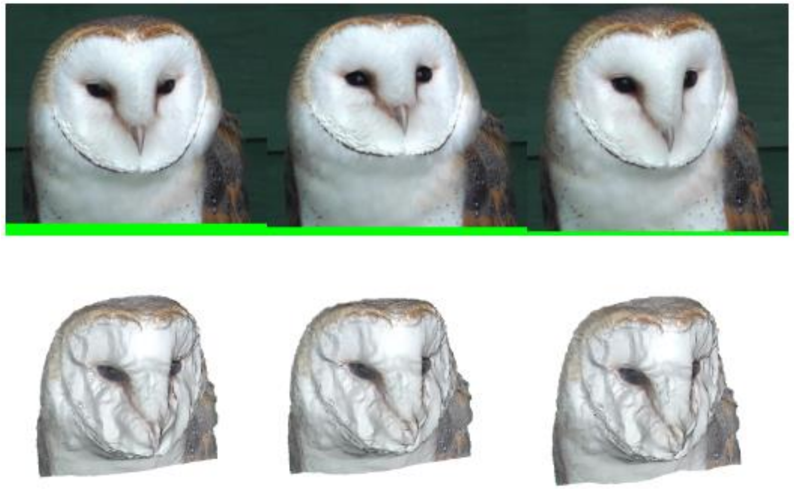


The Sun

Clothes

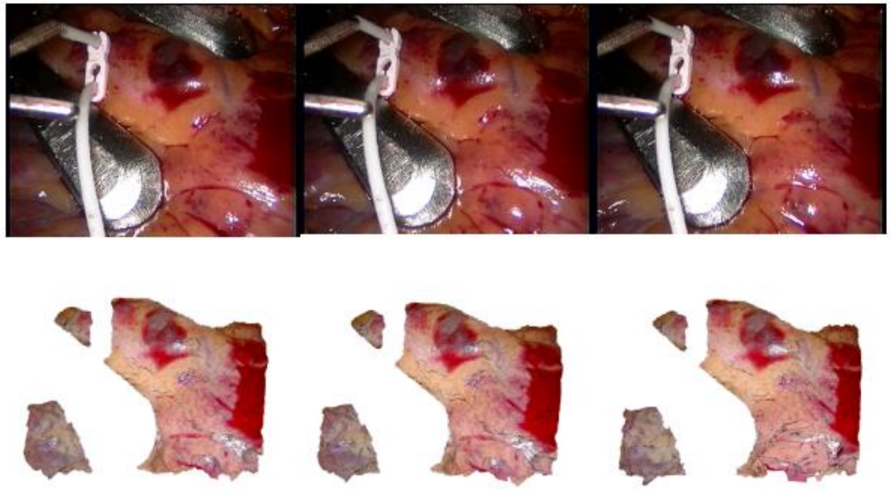


Sheets



Animal face

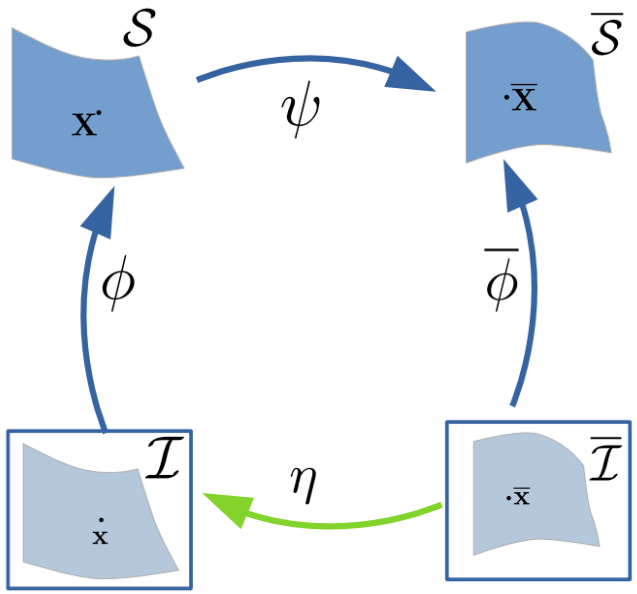
Human back



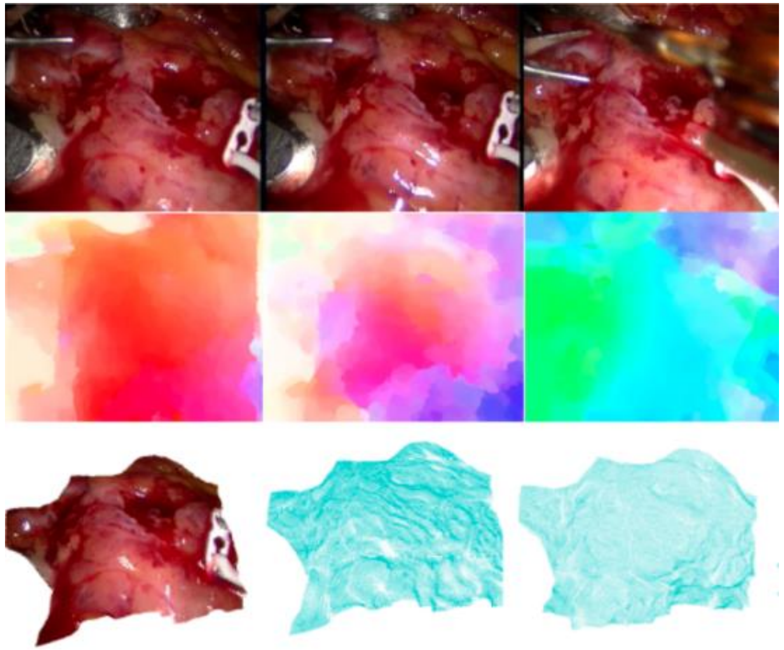
Human body tissues

Dense NRSfM: State of the Art

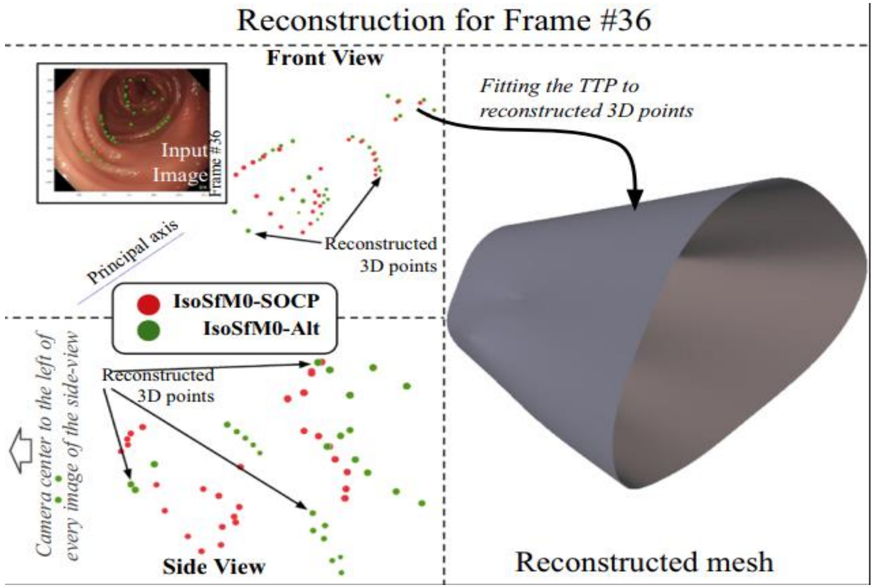
Different priors



Parashar *et al.* 2020 (Local NRSfM from Diffeomorphic Mappings)

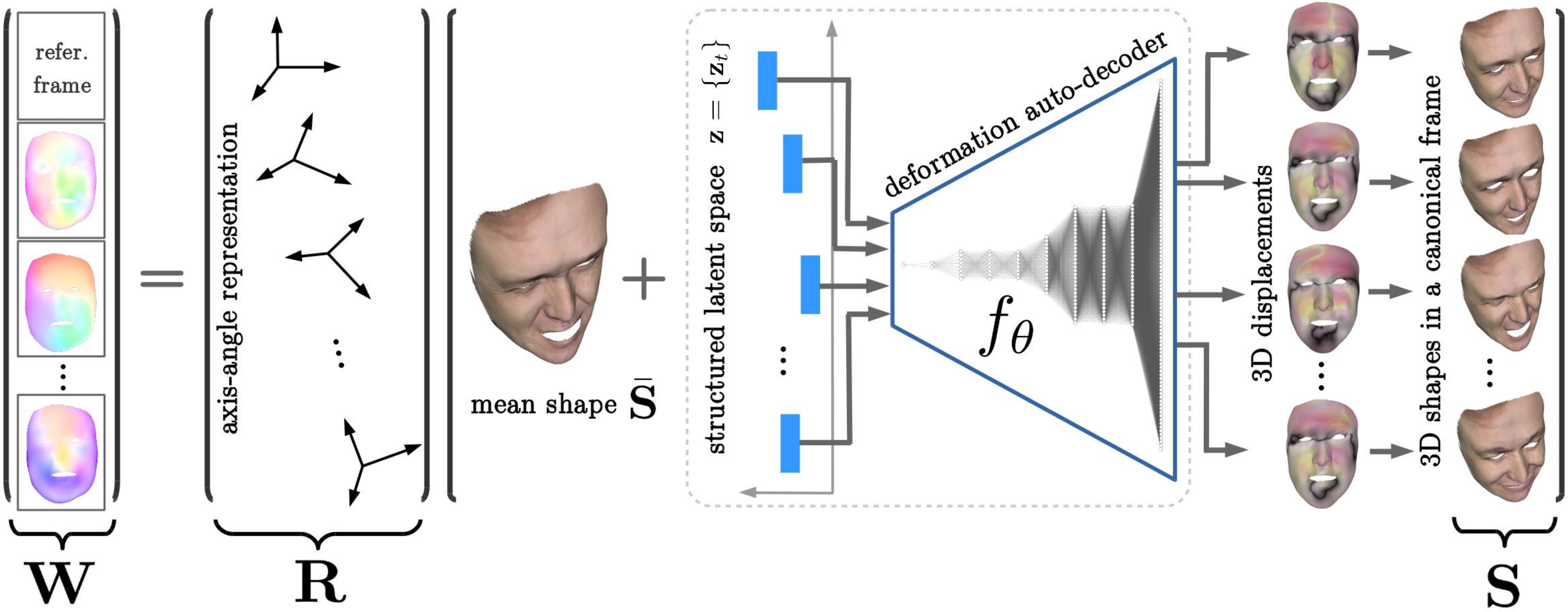


Golyanik *et al.* 2020 (Dynamic Shape Prior)



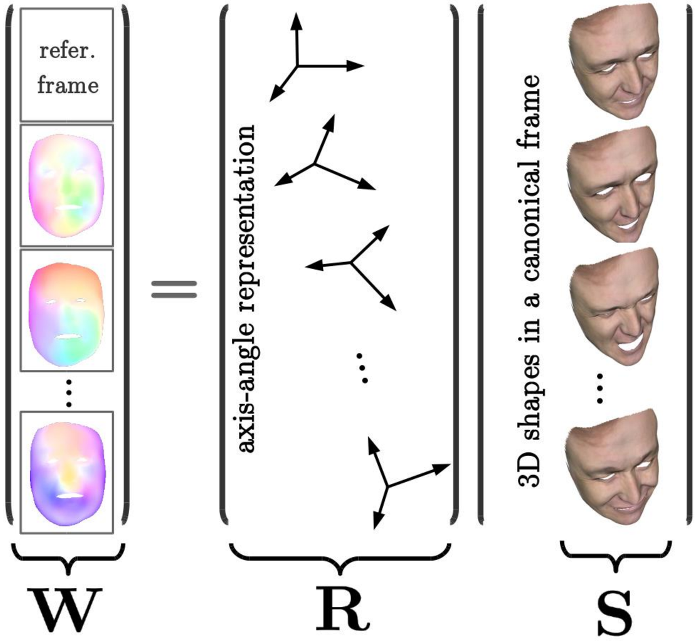
Sengupta *et al.* 2021 (NRSfM with Topological Prior)

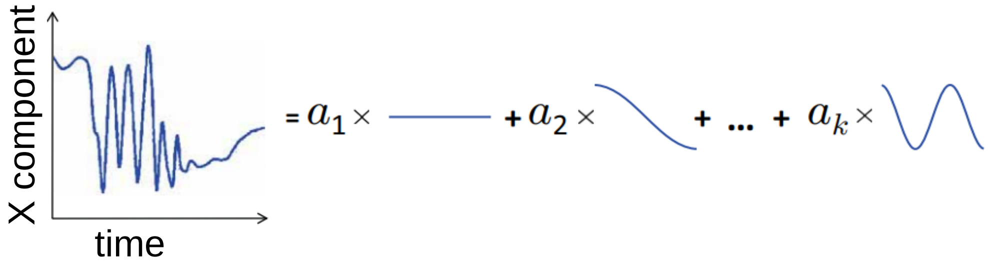
Neural Dense NRSfM

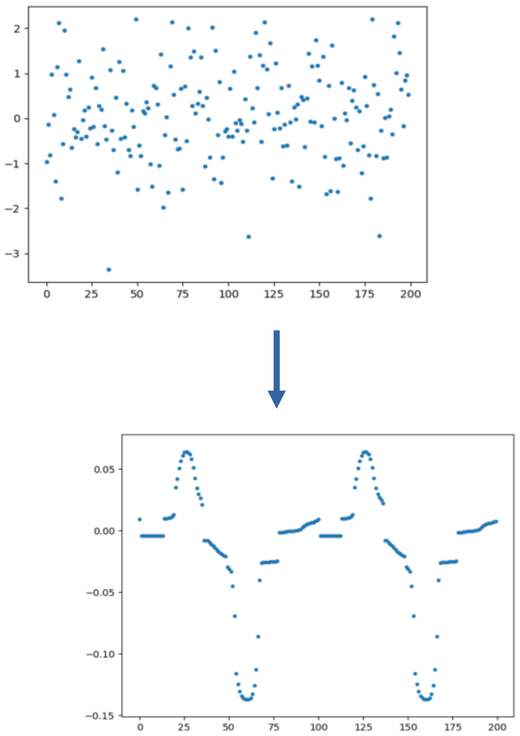


Sidhu *et al.* 2020

Neural Dense NRSfM



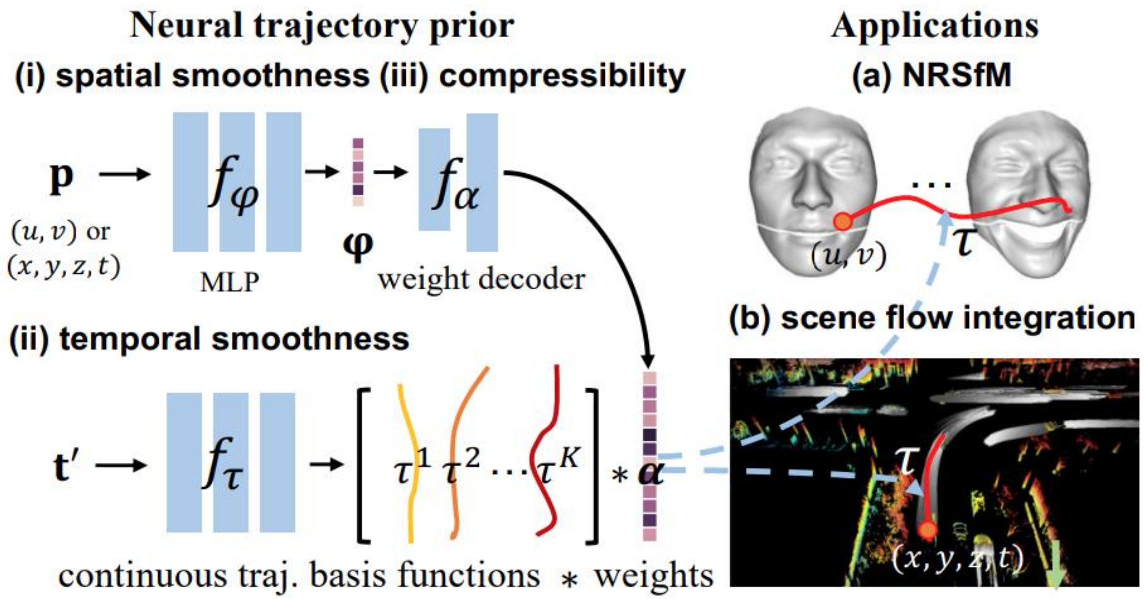


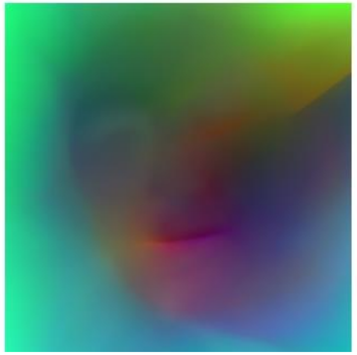


$$\mathbf{E} = \mathbf{E}_{\text{data}}(\boldsymbol{\theta}, \mathbf{z}, \mathbf{R}) + \beta\, \mathbf{E}_{\text{temp}}(\boldsymbol{\theta}, \mathbf{z}) + \gamma\, \mathbf{E}_{\text{spat}}(\boldsymbol{\theta}, \mathbf{z}) + \eta\, \mathbf{E}_{\text{traj}}(\boldsymbol{\theta}, \mathbf{z}) + \omega\, \mathbf{E}_{\text{latent}}(\mathbf{z})$$

Sidhu *et al.* 2020

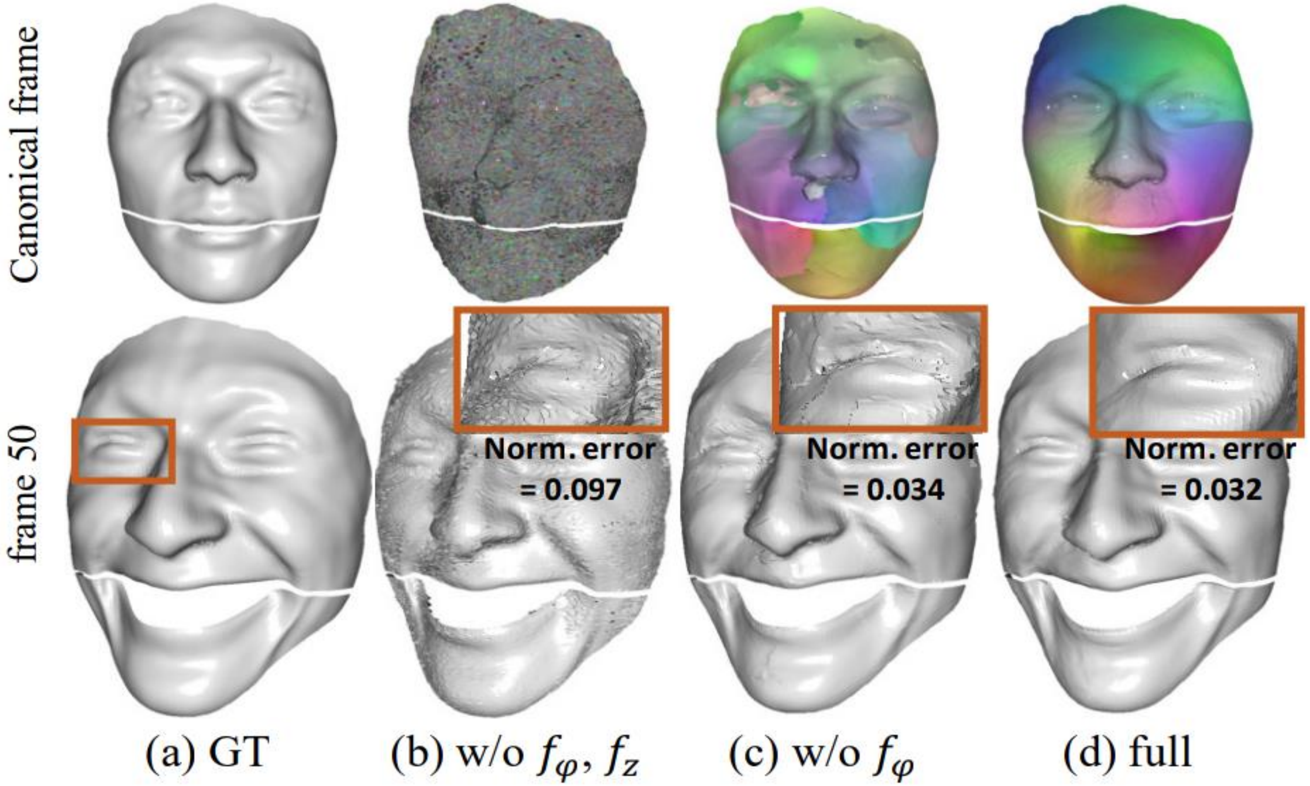
Neural Trajectory Prior for Dense NRSfM





Trajectory
Code Field

Canonical
Reconstruction



Wang *et al*. 2022

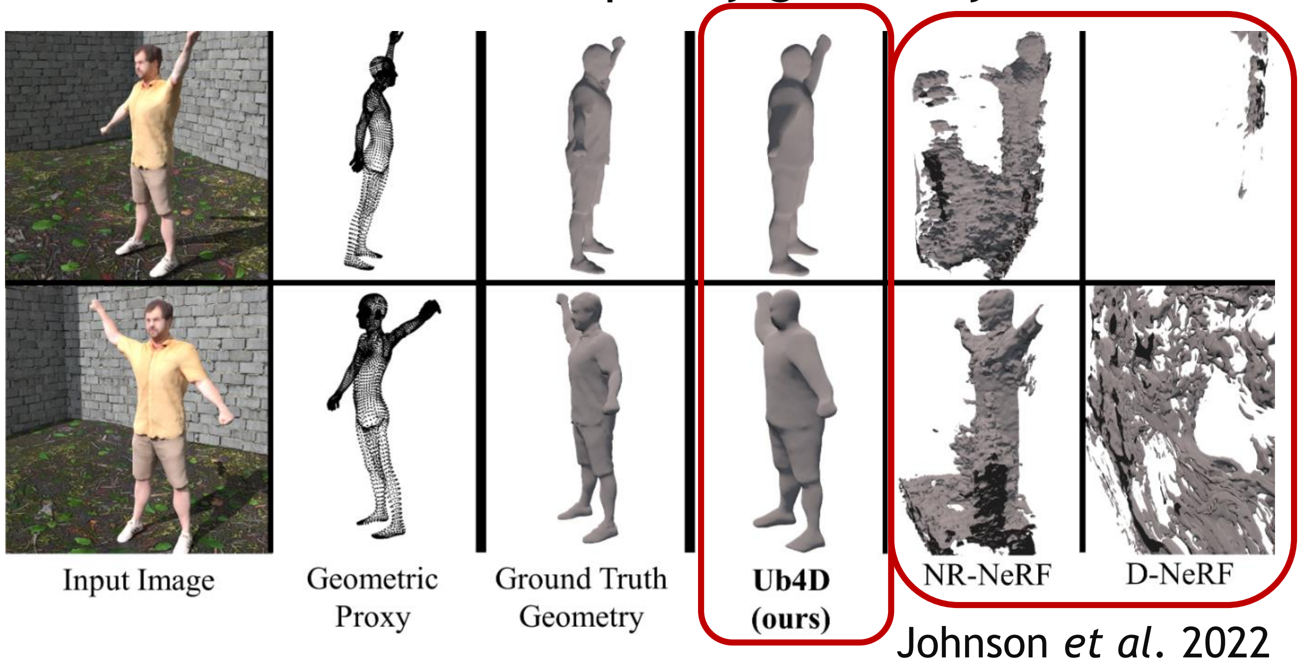
Dense NRSfM: Open Challenges

- NRSfM depends on 2D point tracks → Difficult to obtain
  - Most methods evaluate on ground-truth 2D matches
  - Joint evaluation of 2D flow and regressed 3D shapes is rare
- NRSfM's assumptions (*e.g.* rigidity) are almost never fulfilled in practice
  - Closely related methods (Johnson *et al.* 2023) do not require 2D point tracks or 3D templates
- Saturation in NRSfM:
  - Marginal improvements on existing datasets
  - Small motions
- NRSfM only considers points in first frame → Shape completion remains unsolved

3.1.3 Neural Scene Representations

Neural Scene Representations

- New and very active area besides established SfT and NRSfM
- What do they have in common?
  - Crucially: NeRF-style scene representation and volumetric rendering
  - No template → Also reconstruct background
  - Focus on novel view synthesis → Density function → Rather low-quality geometry
    - But: Better geometry (Johnson *et al.* 2023)



Johnson *et al.* 2022

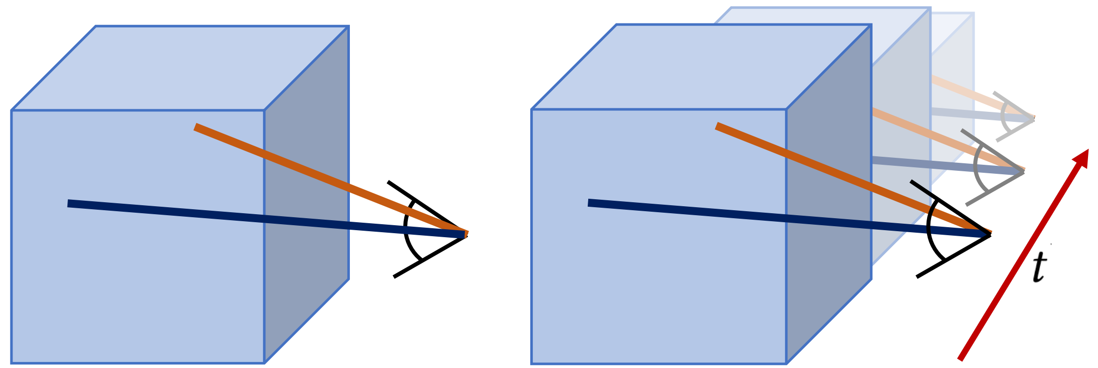
Neural Scene Representations

- New and very active area besides established SfT and NRSfM
- What do they have in common?
  - Crucially: NeRF-style scene representation and volumetric rendering
  - No template → Also reconstruct background
  - Focus on novel view synthesis → Density function → Rather low-quality geometry
    - But: Better geometry (Johnson *et al.* 2023)
  - Slow: Many hours to reconstruct one scene
    - But: Recent methods only take a few minutes (Fang *et al.* 2022, Guo *et al.* 2022)
- Lots of different input annotations, *e.g.* camera parameters, optical flow, segmentation masks, static background points
- No standard datasets, mostly self-recorded videos (see also Gao *et al.* 2022)

How to Parametrize Deformations

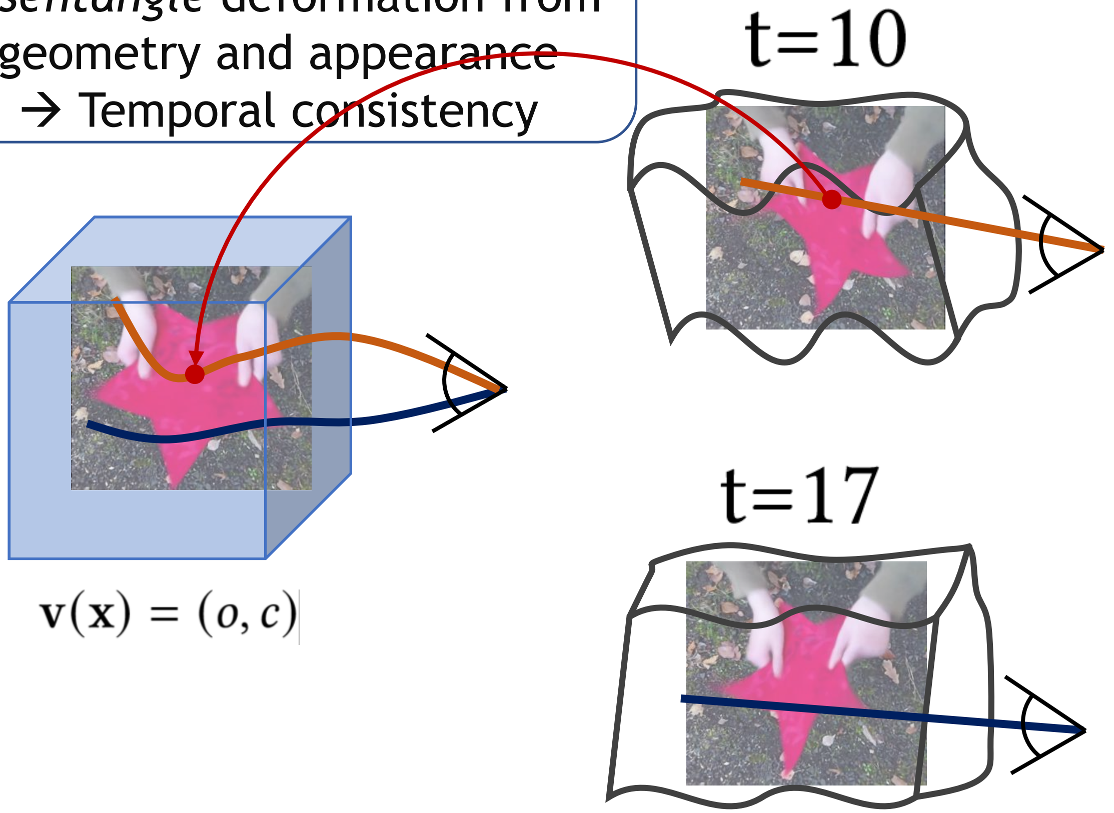
Time Conditioning:
Entangle deformation with geometry and appearance
→ Challenging motion

Ray Bending:
Disentangle deformation from geometry and appearance
→ Temporal consistency



$\mathbf{v}(\mathbf{x}) = (o, c)$

$\mathbf{v}(\mathbf{x}, t) = (o, c)$



$\mathbf{v}(\mathbf{x}) = (o, c)$

- Trade off between challenging motion and long-term temporal consistency
  - Little progress in terms of reconstruction quality, rather shifting of trade off

3.1.4 Other Few-Scene Methods

Other Methods for Few Scenes

- Methods that:
  - Do not fall into the previous categories
  - And reconstruct a single or a few scenes
  - → Parametrizing each scene directly is still feasible

Yang *et al.* 2022

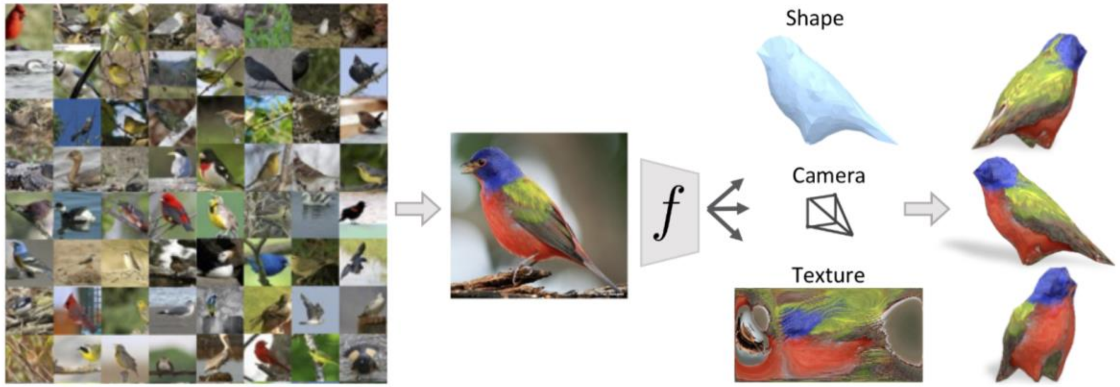
Method	Geometry	Correspondences	Number of Scenes
Yang *et al.* 2021: LASR	Mesh	RGB appearance	One video
Yang *et al.* 2021: ViSER	Mesh	RGB + learned features	A few videos
Yang *et al.* 2022: BANMo	NeRF	Pretrained features + RGB	A few videos
Yao *et al.* 2022: LASSIE	Mesh	Pretrained features	Ca. 30 images

- Common themes:
  - Differentiable rendering: Naturally connects 2D input and 3D reconstruction
  - Learned features: Robustness to appearance variations (*e.g.* from the environment, deformations, multiple individuals)
  - Neural representations: Easier optimization than meshes

3.1.5
Learned Prior

Data-Driven Priors

- Became possible due to deep learning and differentiable mesh rendering
- Training: Learn a prior from an image collection of many scenes
- Test: *Regress* scene parameters of an *unseen* scene



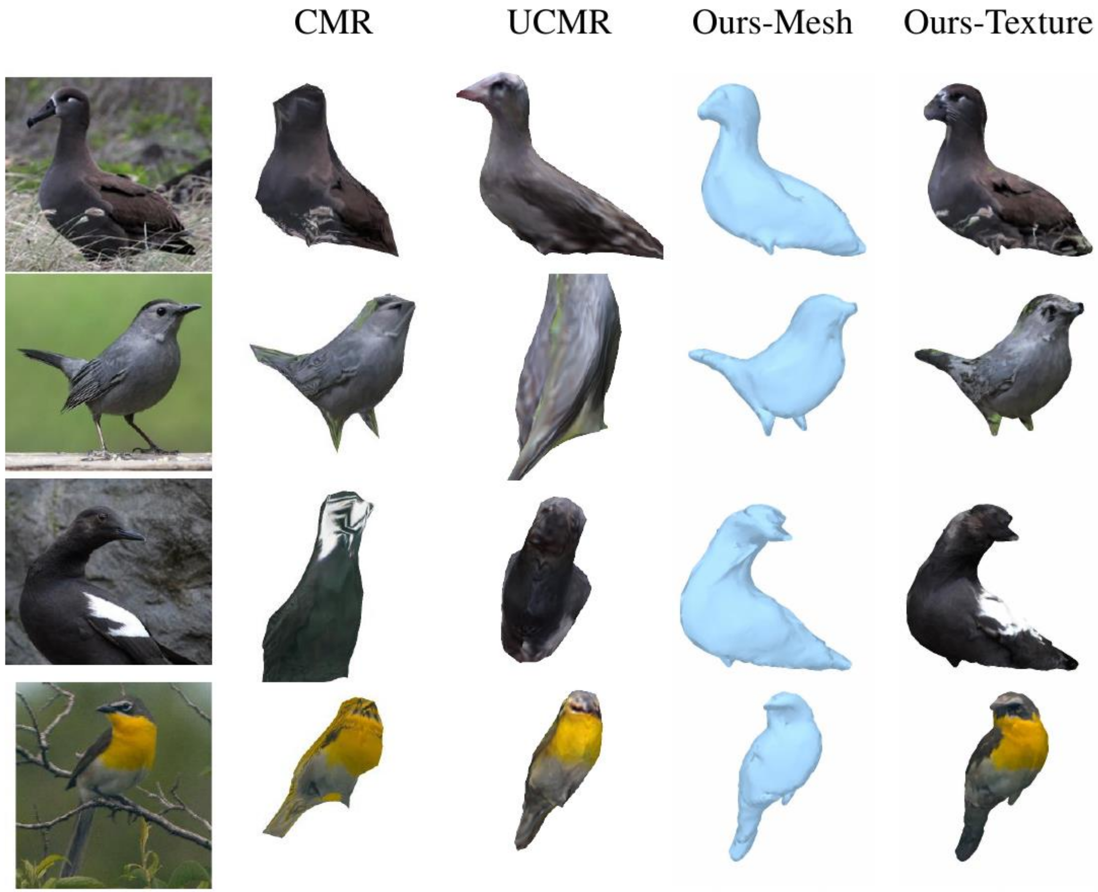
Kanazawa *et al*. 2018 (CMR)

- General trends:
  - Focus on CUB dataset of birds (Wah *et al*. 2011)
  - Barely any qualitative improvement over a dozen papers:
    - Appearance: Fairly detailed (by sampling from the input image)
    - Geometry: Very coarse, *e.g*. wings or legs are still hardly reconstructed
  - Reduce input annotations, explore alternative inputs like videos

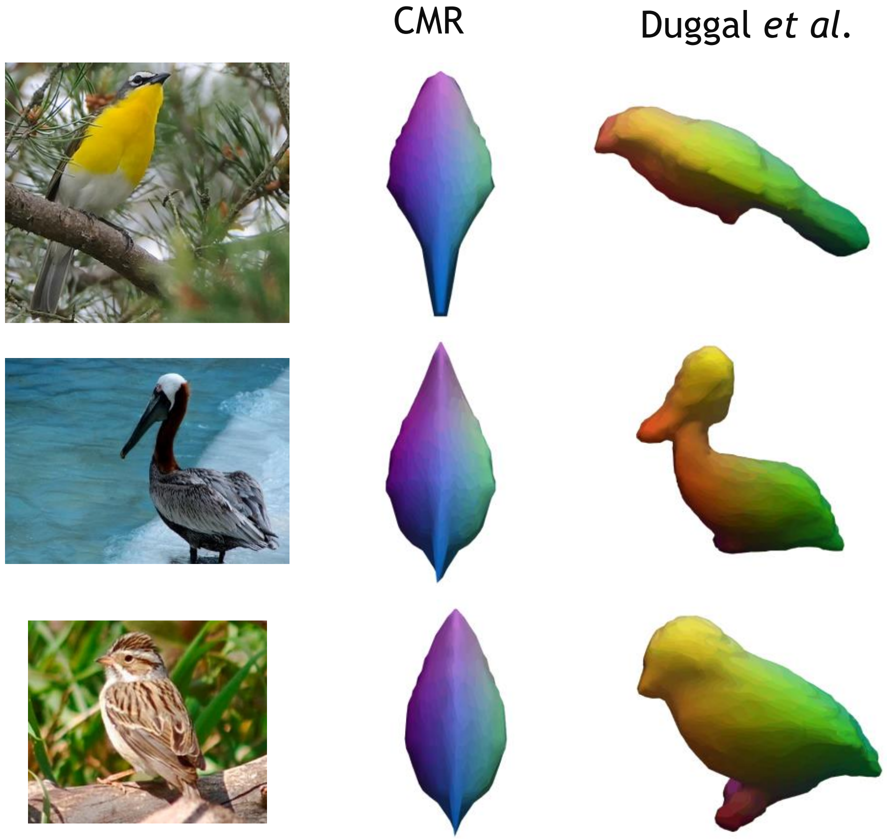
Data-Driven Priors

- Recently: Noticeable improvements by allowing more variation from the template

Kokkinos *et al.* 2021:
At training time, regress + *refine* deformations

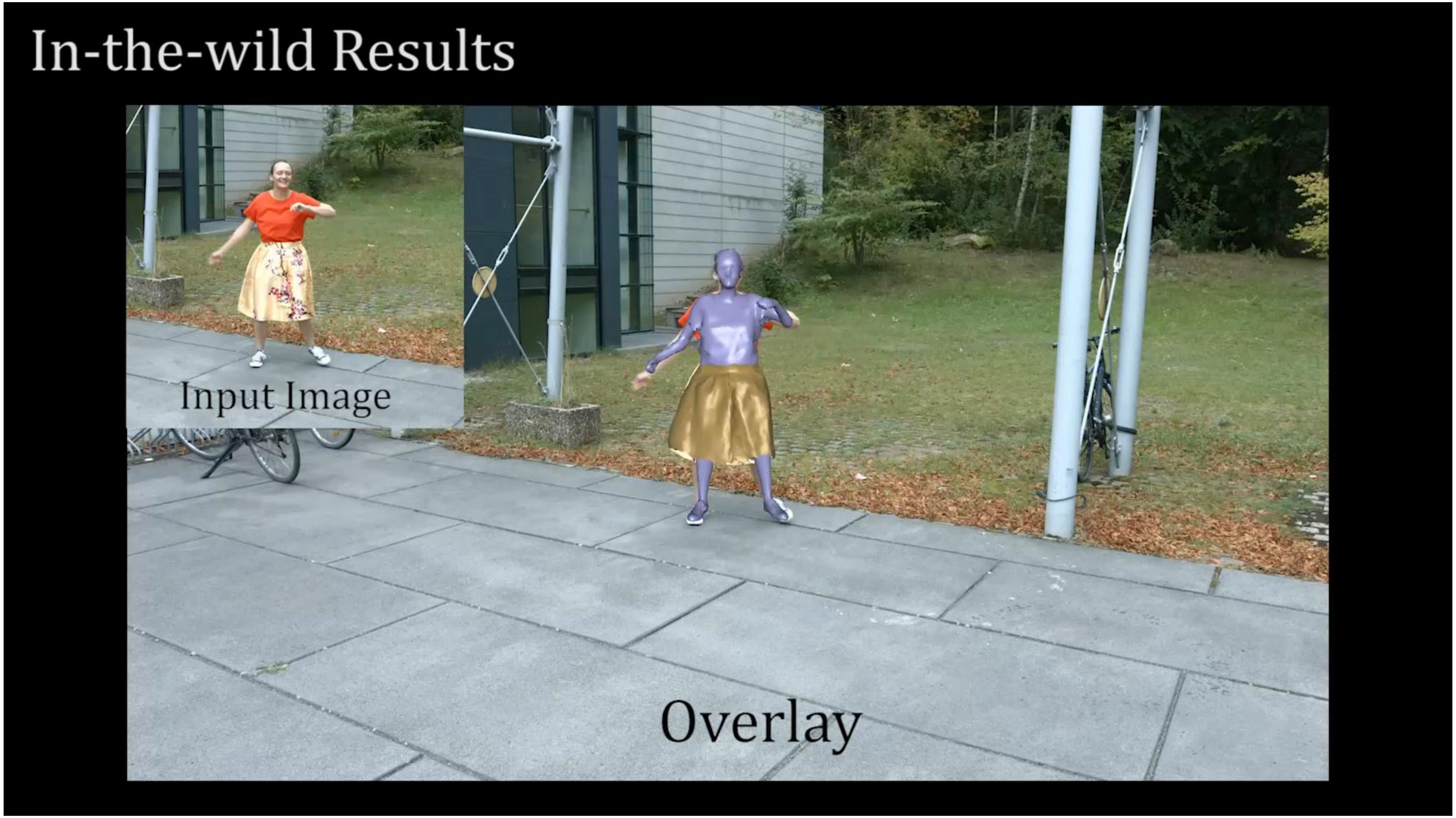


Duggal *et al.* 2022:
Neural representation + regress template for each image



3.2 Humans

Dense Monocular Human Reconstruction



Li *et al.* 2021

Challenges

Large Displacements

Loose Clothing

Self-Occlusions

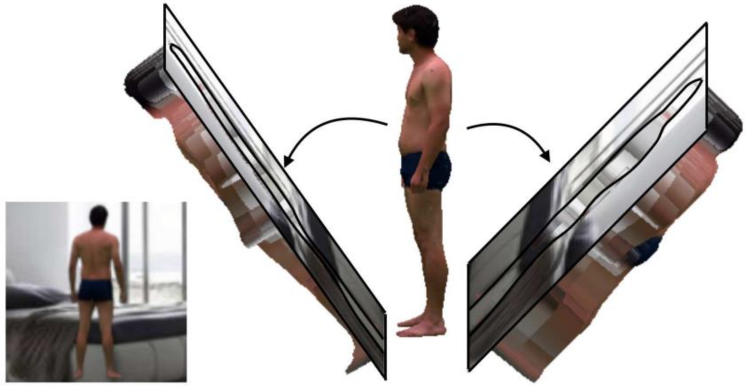
He *et al.* 2022

Challenges

Piecewise rigid deformations + Non-rigid surface deformations

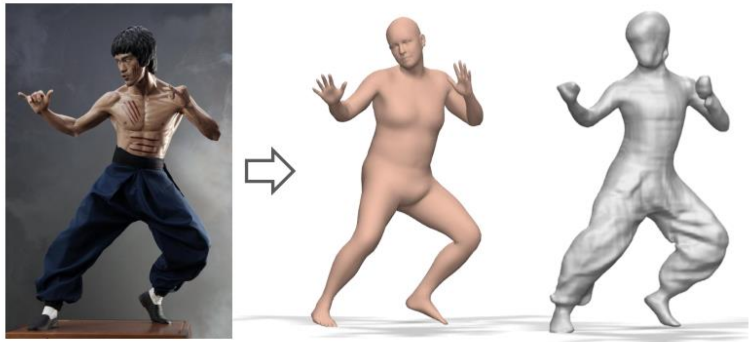
Taxonomy

Template-Free



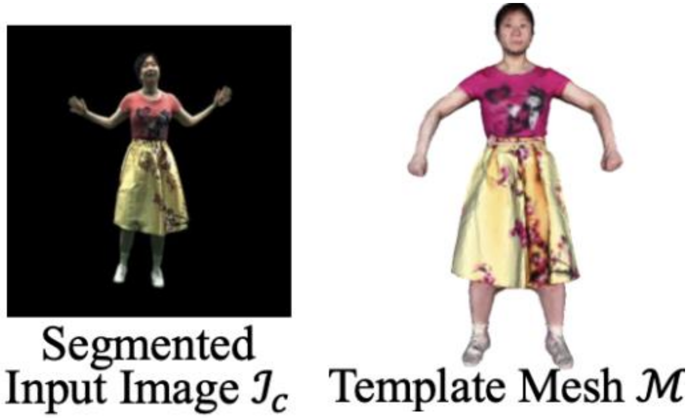
Gabeur *et al.* 2019

Parametric Models (SMPL, GHUM, etc.)



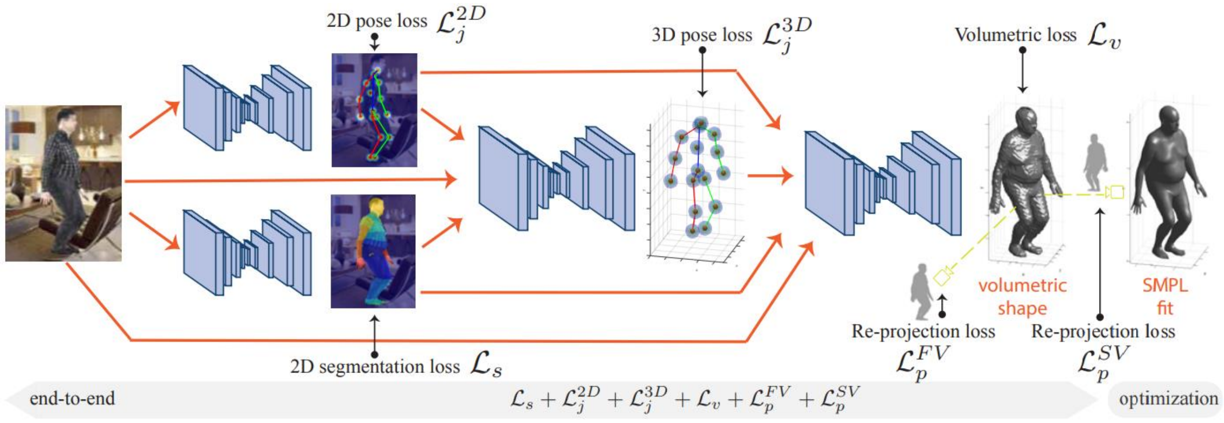
Zheng *et al.* 2021

Template-Based

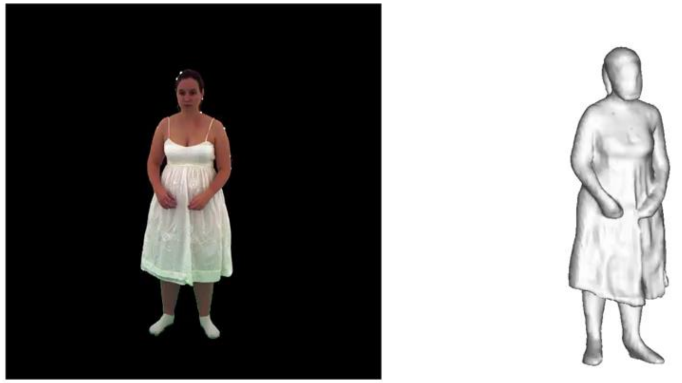


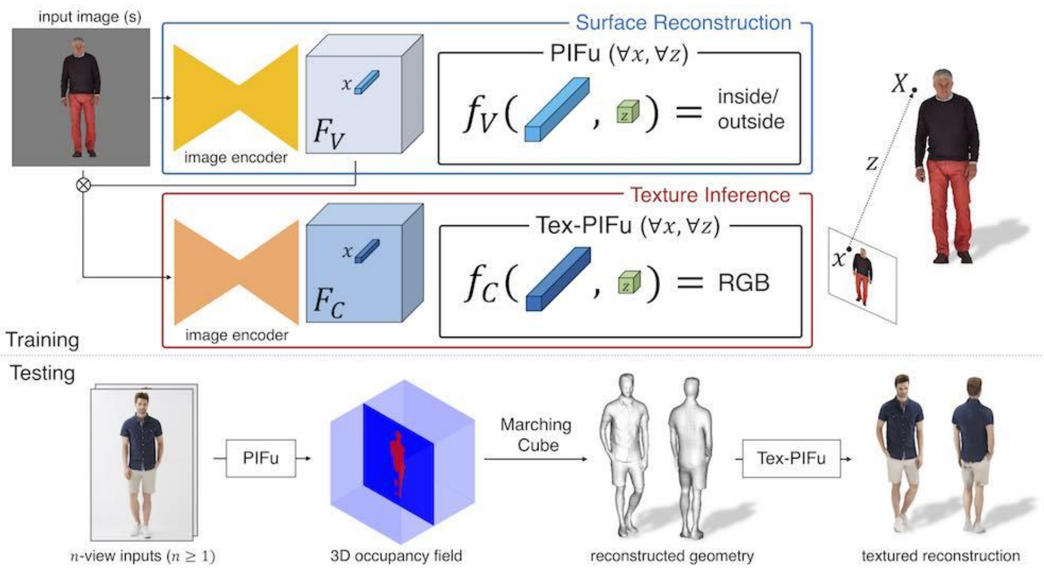
Jiang *et al.* 2022

Template-Free Methods



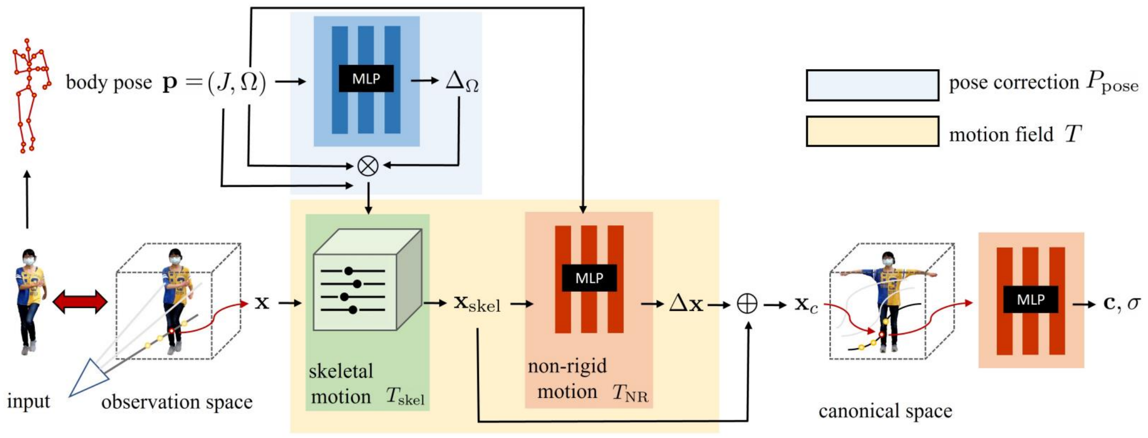
Varol *et al.* 2018 (BodyNet)





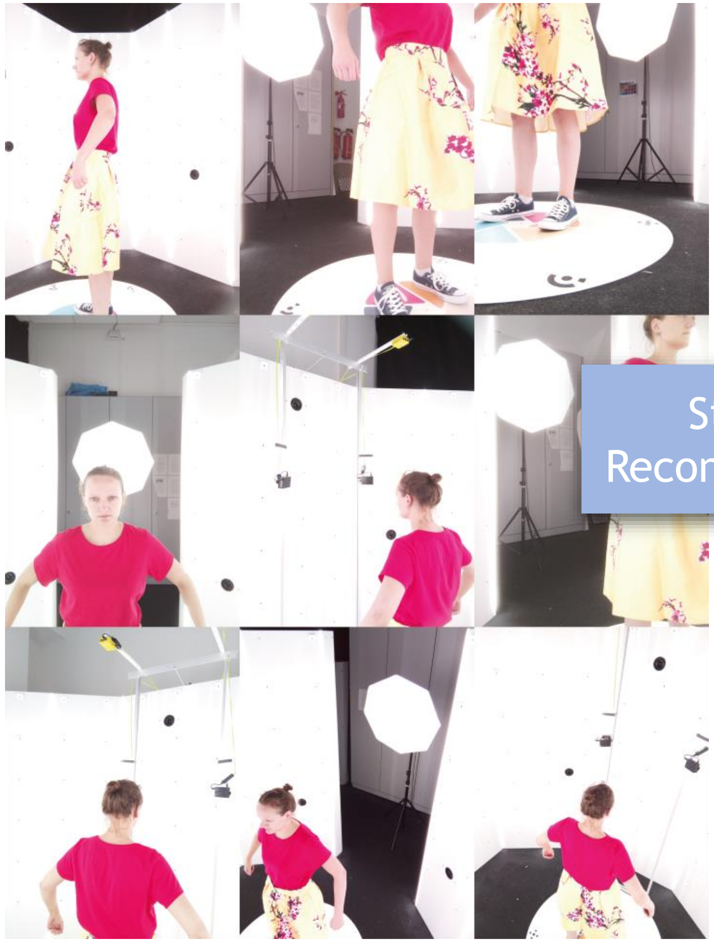
Saito *et al.* 2019 (PIFu)

Template-Free Methods



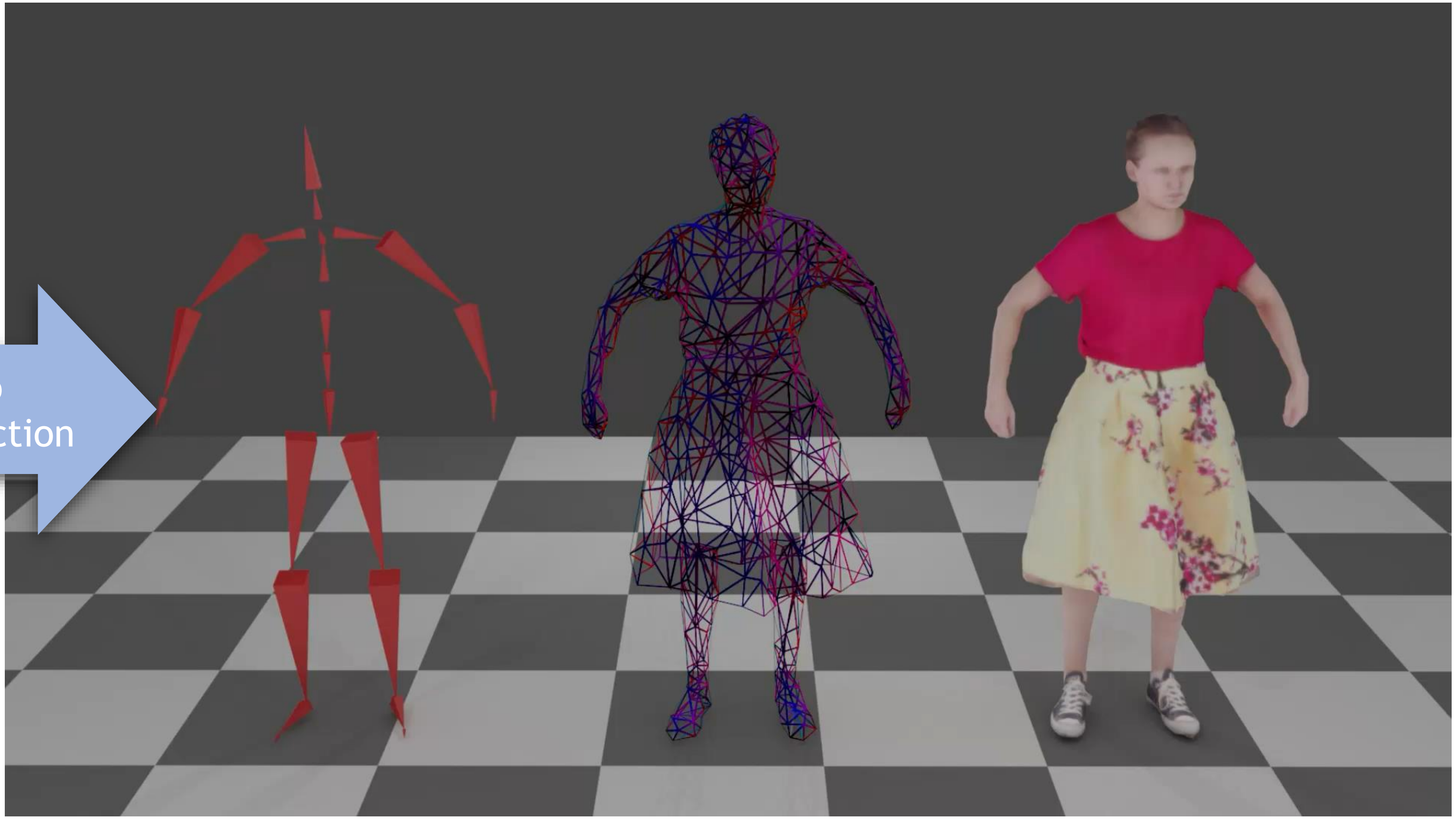
Weng *et al.* 2022 (Human-NeRF)

Template-Based Methods



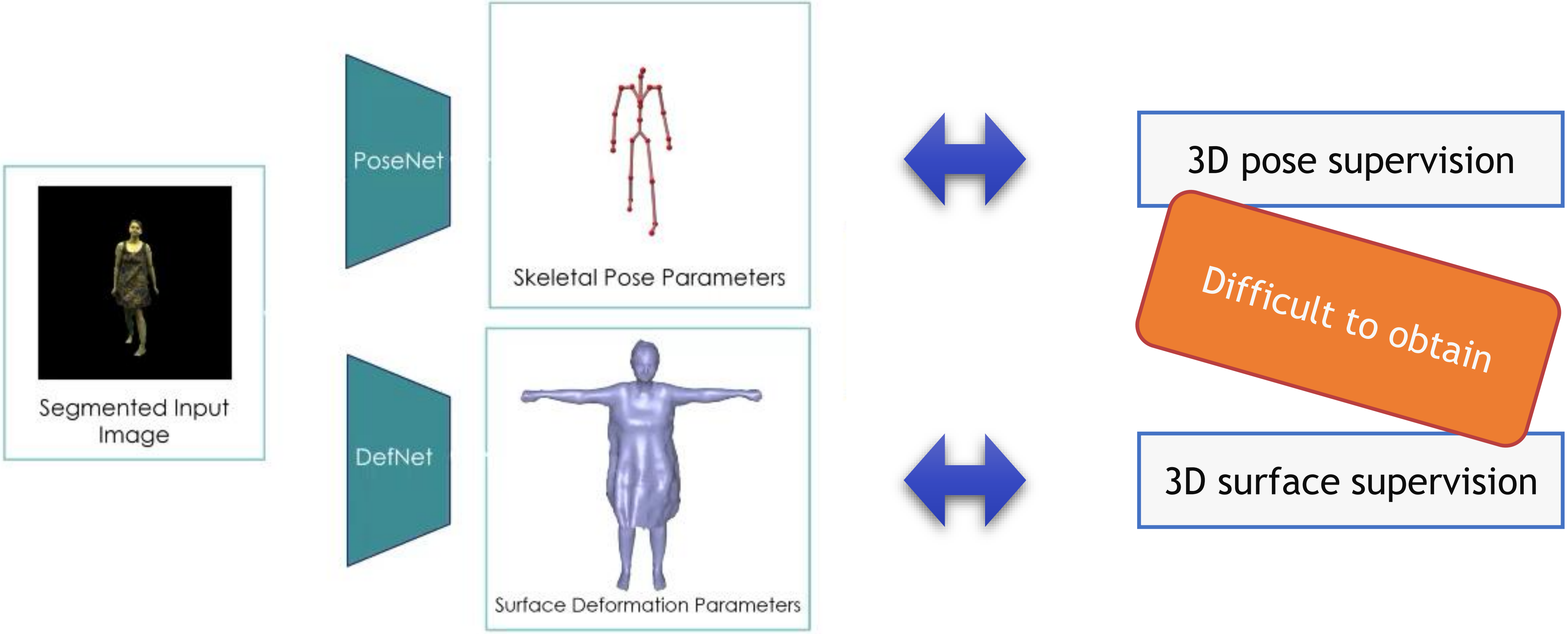
Multi-View Images

Stereo
Reconstruction



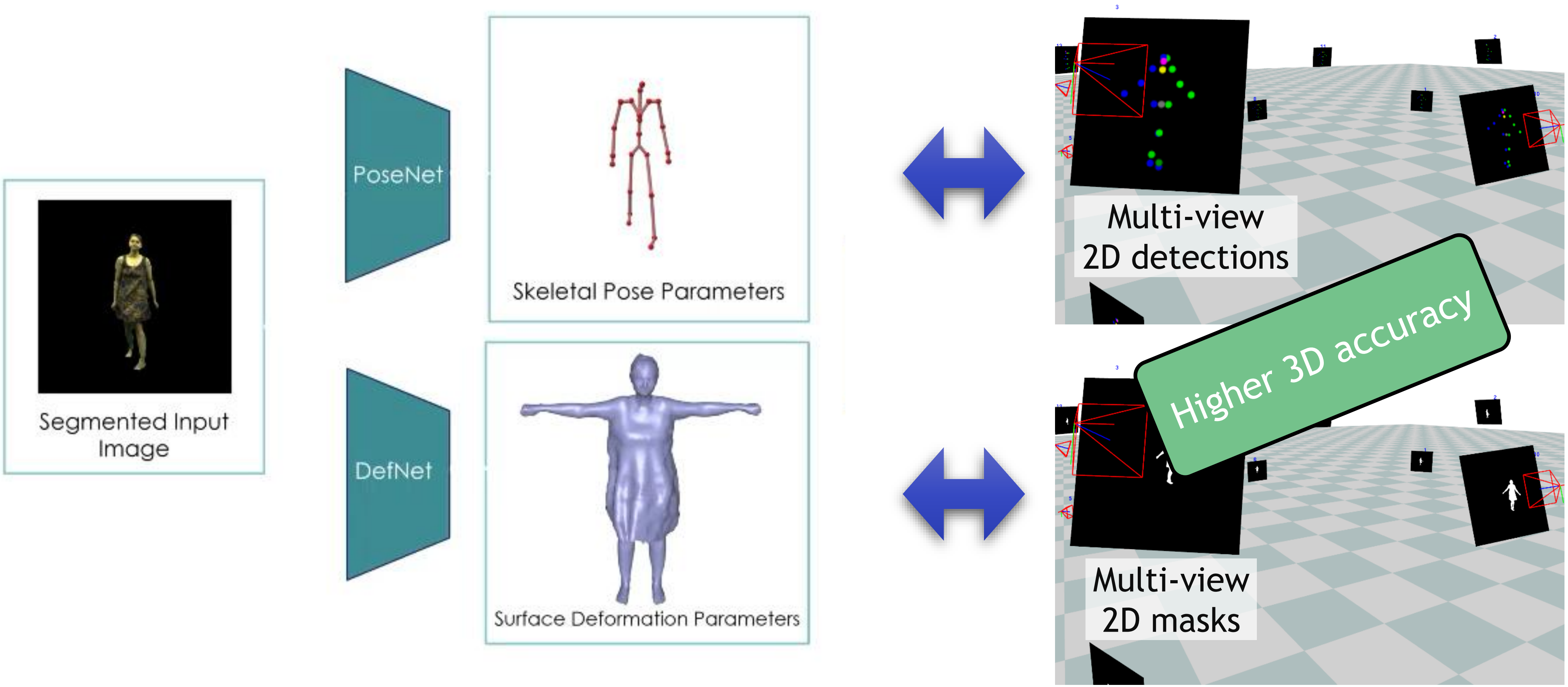
Skeleton Embedded Graph Textured Template Mesh

DeepCap



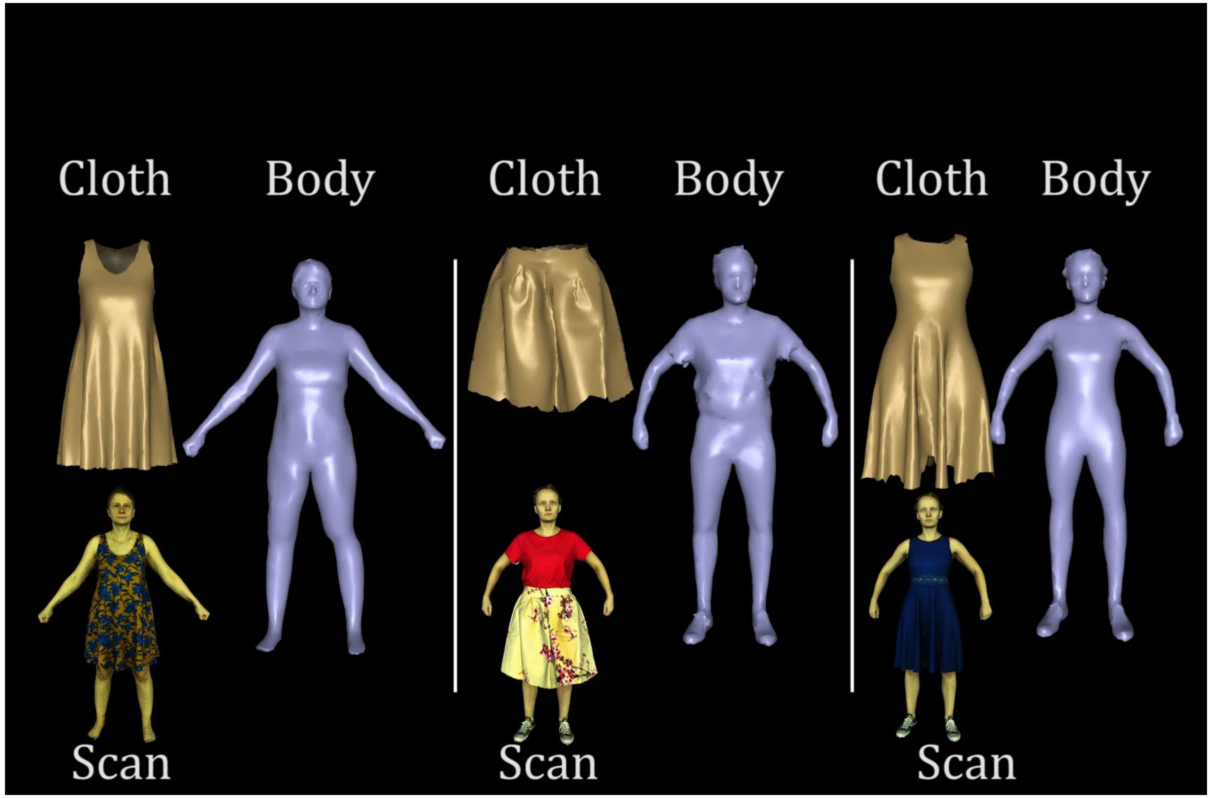
Habermann *et al*. 2020

DeepCap

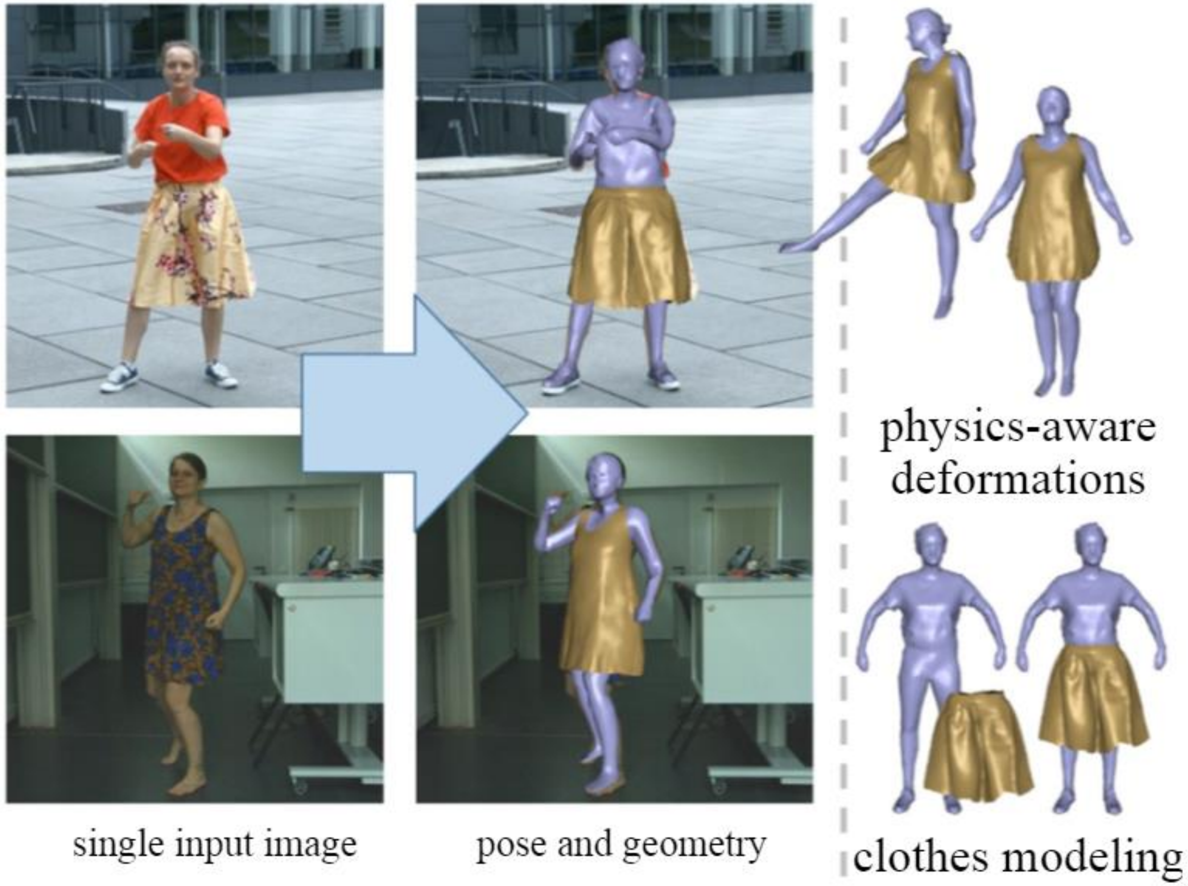


Habermann *et al*. 2020

Introducing Cloth Physics

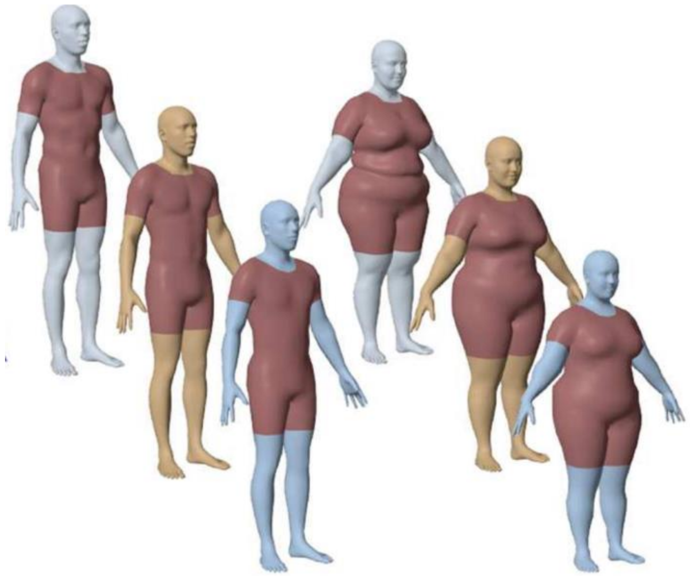


Updated Template Mesh

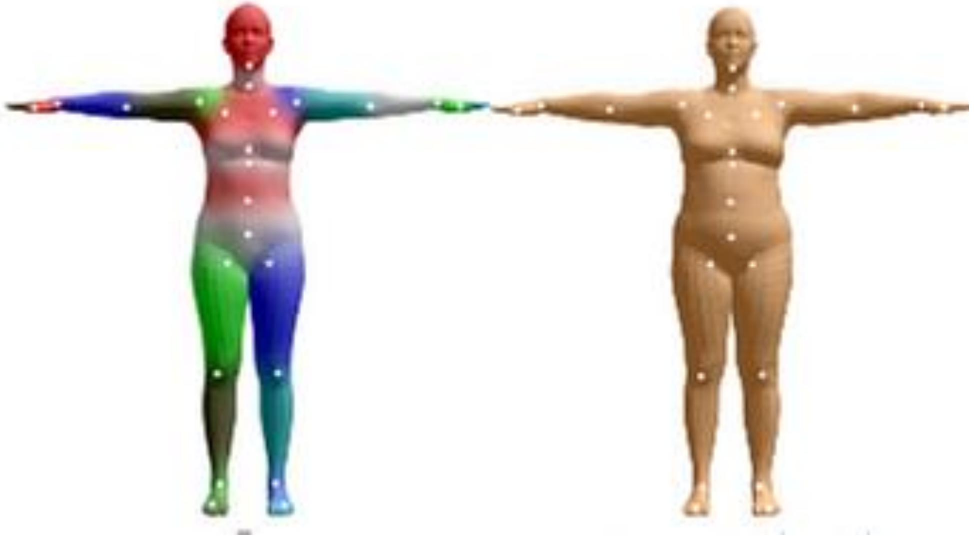


Li *et al.* 2021

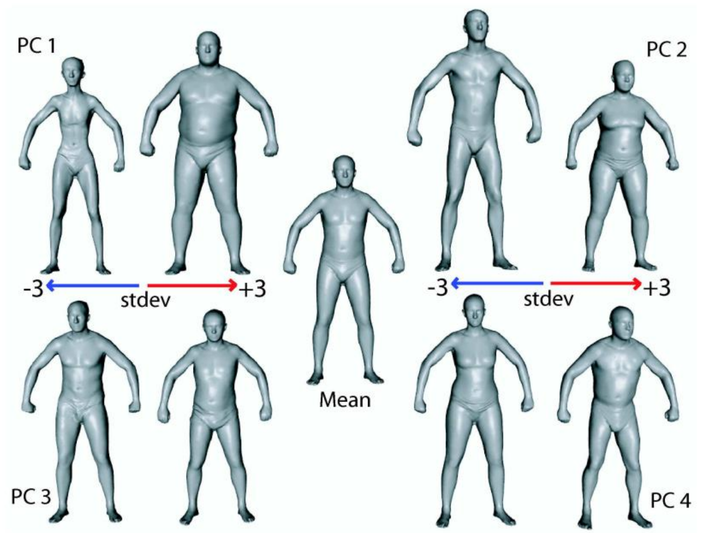
Using Parametric Models



GHUM(L)
(Xu *et al.* 2020)



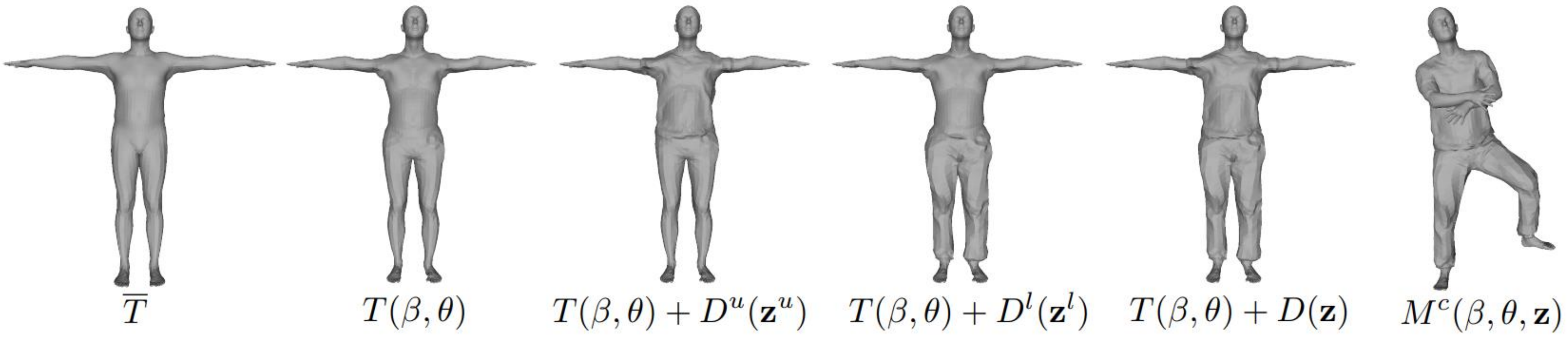
SMPL
(Loper *et al.* 2015)

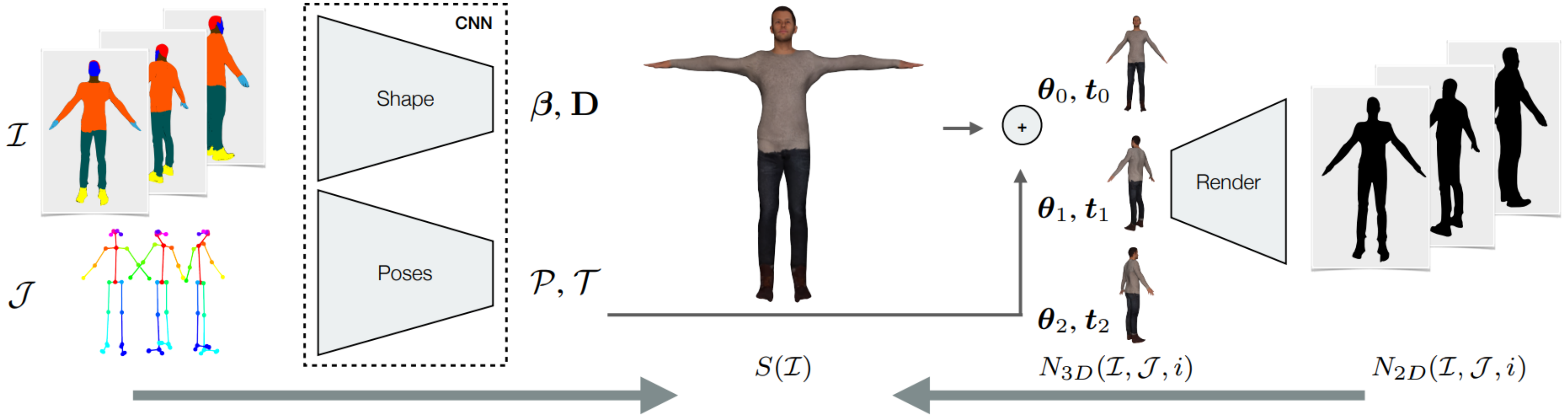


SCAPE
(Anguelov *et al.* 2005)

Deforming the Parametric Models

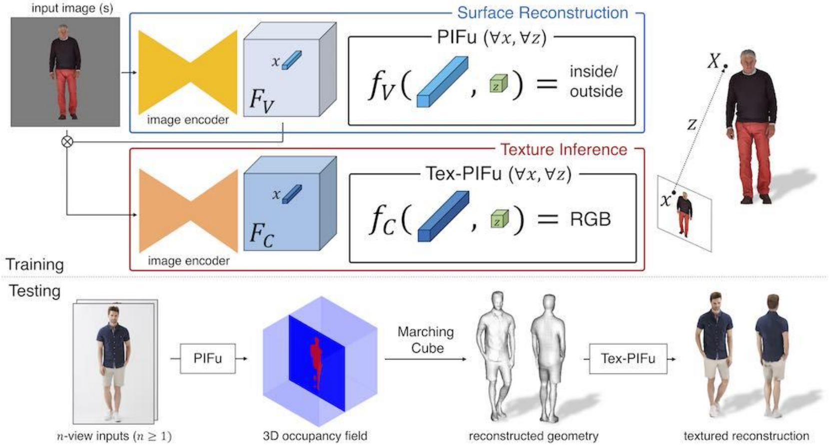
Xiang *et al.* 2020 (MonoClothCap)





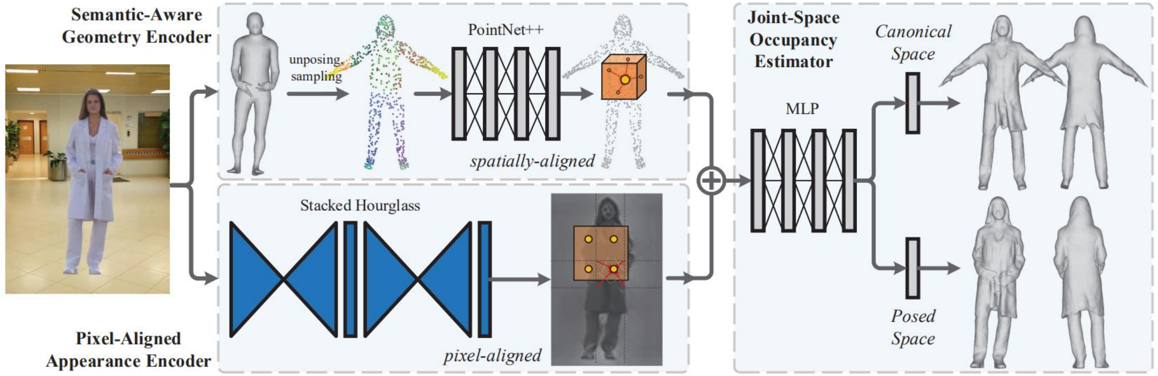
Alldieck *et al.* 2019

Parametric Models as Geometric Priors



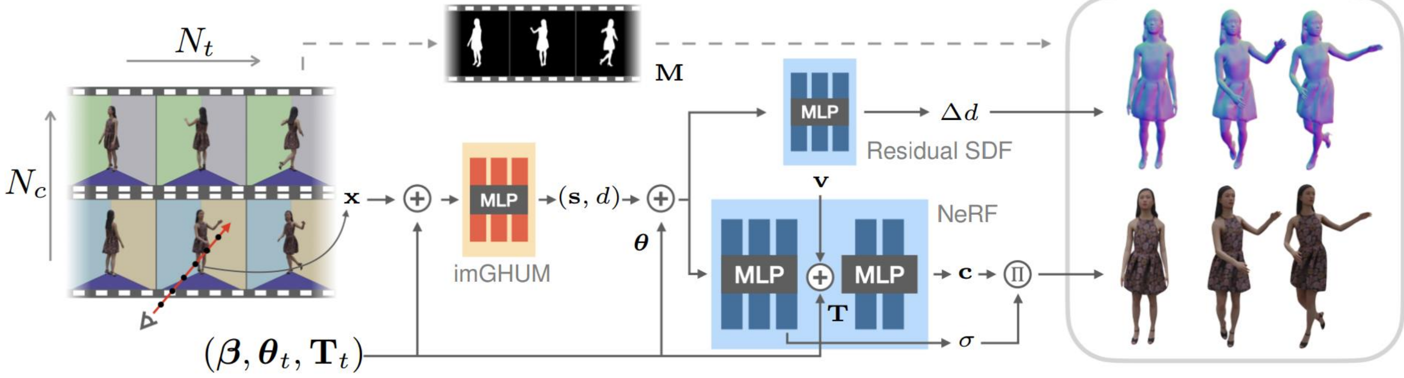
Saito *et al.* 2019 (PIFu)

Template-Free!

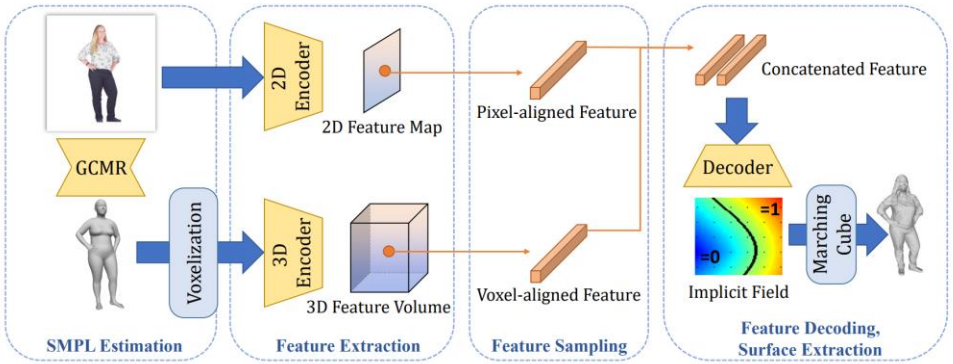


He *et al.* 2021 (ARCH++)

Parametric models provide strong *geometric* features

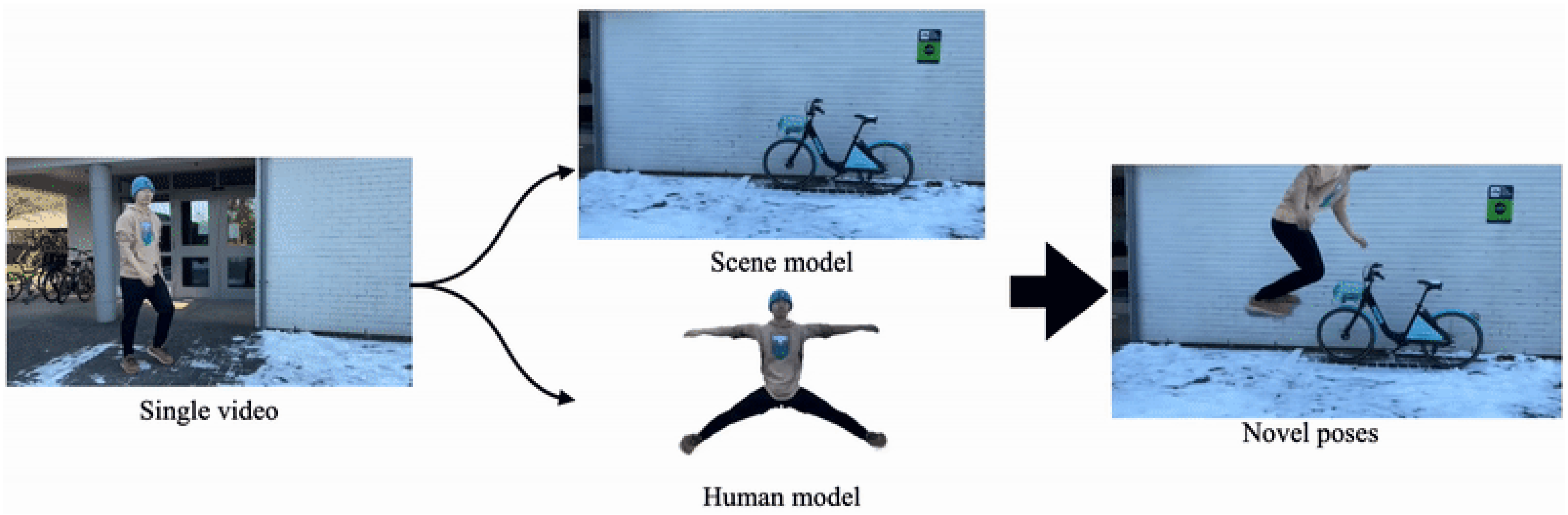


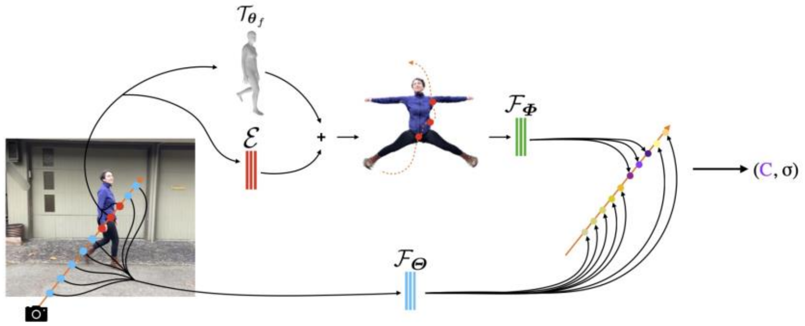
Xu *et al.* 2022 (H-NeRF)



Zheng *et al.* 2021 (PaMIR)

Joint Human-Scene Reconstruction





Wei *et al.* 2022 (NeuMan)

Future Directions

- Parametric models for geometry *and appearance*
- Tracking of topological changes
- Joint dense body capture (including hands, face, gaze, hair, etc.)
- Robustness and interpretability

Zhu *et al*. 2020 (DeepFashion3D)

3.3 Faces

Faces

Input Reconstruction Geometry

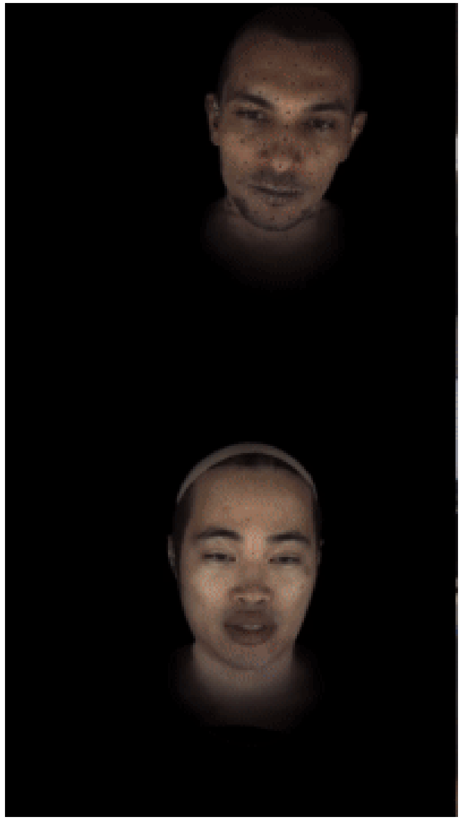
Input Reconstruction

What’s Special About Faces?

- Easier to build priors!
  - Relatively (human body) not much articulation
  - Regular pattern: Symmetry, fixed parts, etc.
  - Less diversity: No clothing, less accessories, etc.
  - Availability of large-scale data

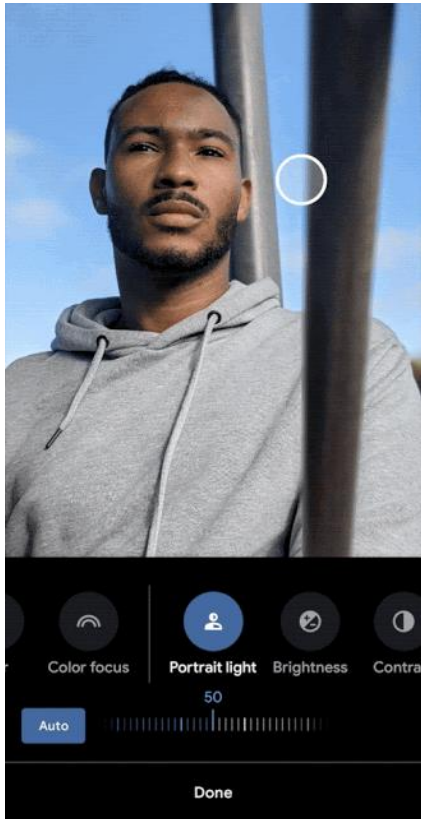
- Challenges:
  - Hair has complex geometry and topology
  - Even minor misprediction could lead to perceptually significant difference

Applications



VR/AR
©Meta

Movies / Gaming
©Weta



Mobile Applications
©Google

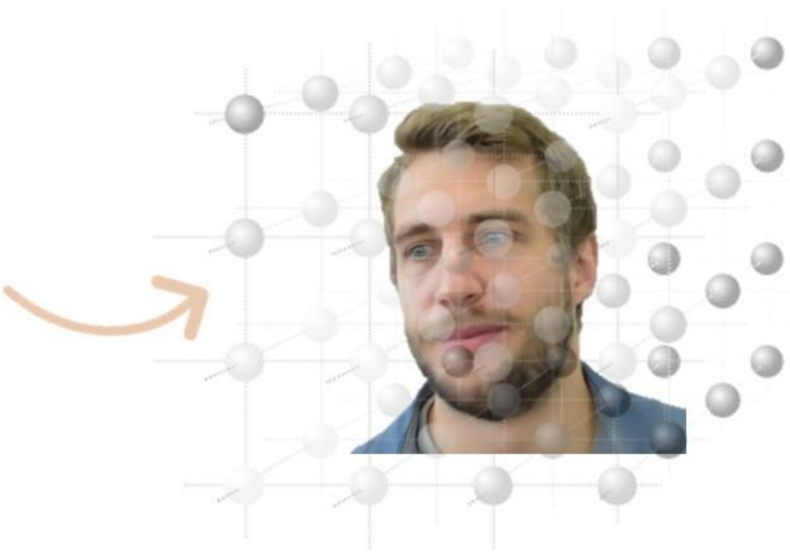
Problem Statement

Single Image

Chan *et al.* 2021 (EG3D)

Li *et al.* 2023 (FOCUS)

Monocular Video



Controllable Avatar
Gafni *et al.* 2021 (NerFACE)

Categorization

Explicit Morphable Models

- Mesh-based representation
- Fixed resolution and topology

Pros:
- Gives SOTA on current benchmarks (at least for the non-hair region)
- Extensively researched

Cons:
- Hard to model thin structures and varying topology, *e.g.* hair

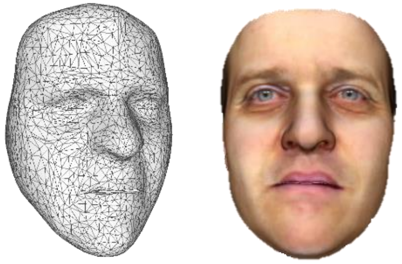
Implicit Morphable Models

- Continuous representation
- Can represent any topology with unlimited resolution

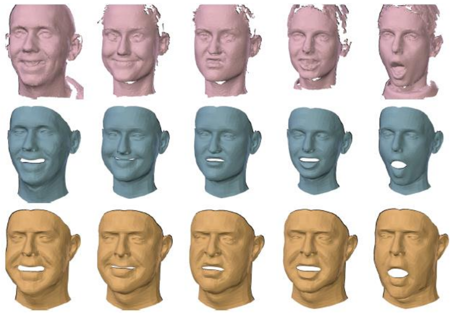
Pros:
- Can model complex geometry, *e.g.* hair
- Easy to model and learn from large-scale data

Cons:
- Not as efficient as explicit models

Explicit Morphable Models



Data Representation



3D Scans

- **Additive model**
  - PCA: Blanz and Vetter 1999, …
  - Blendshapes: Garrido *et al*. 2013, Wu *et al*. 2016, Thies *et al*. 2016, …
- **Multilinear model**
  - Vlasic *et al*. 2005, Cao *et al*. 2014, Shi *et al*. 2014, …
- **Nonlinear model**
  - Li *et al*. 2017, Ichim *et al*. 2017, Shin *et al*. 2014, …

Model Type

Recent Surveys

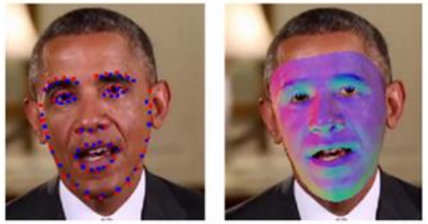
- Zollhöfer *et al*. 2018 (State of the Art on Monocular 3D Face Reconstruction, Tracking, and Applications)
- Egger *et al*. 2020 (3D Morphable Face Models -- Past, Present and Future)

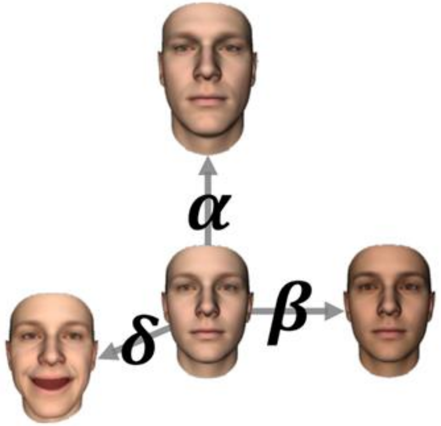
Explicit Morphable Models: Fitting

$$\boldsymbol{E}(\boldsymbol{P}) = \underbrace{w_{col}E_{col}(P) + w_{lan}E_{lan}(P)}_{\boldsymbol{data}} + \underbrace{w_{reg}E_{reg}(P)}_{\boldsymbol{prior}}$$

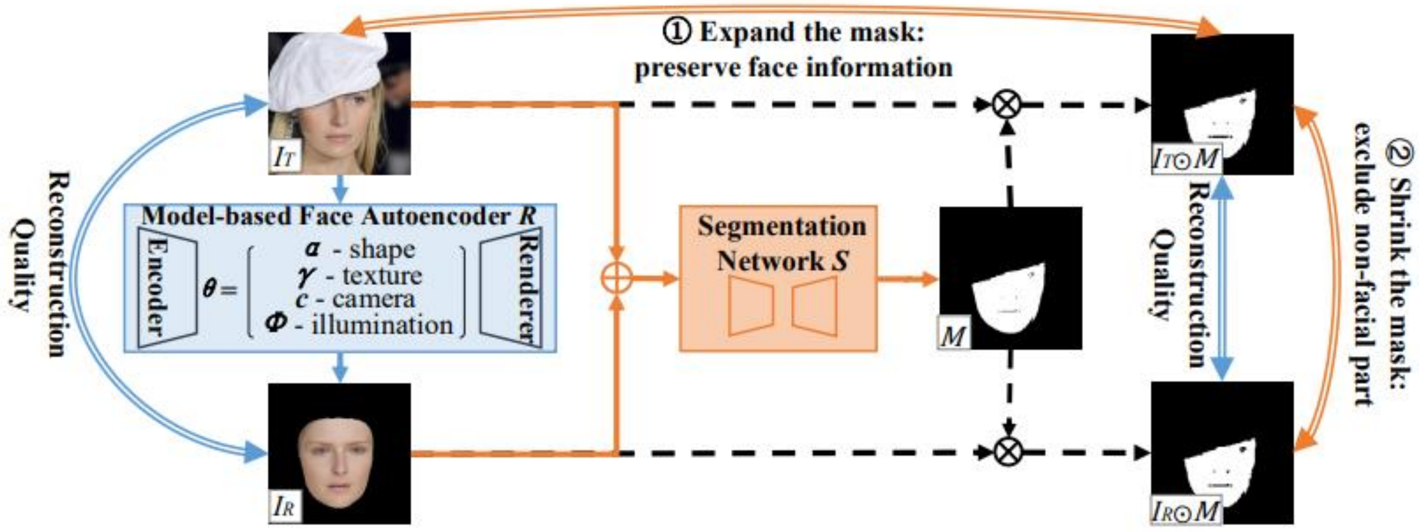
$$\boldsymbol{P} = (\boldsymbol{\Phi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta})$$

pose
shape
albedo
illumination
expression



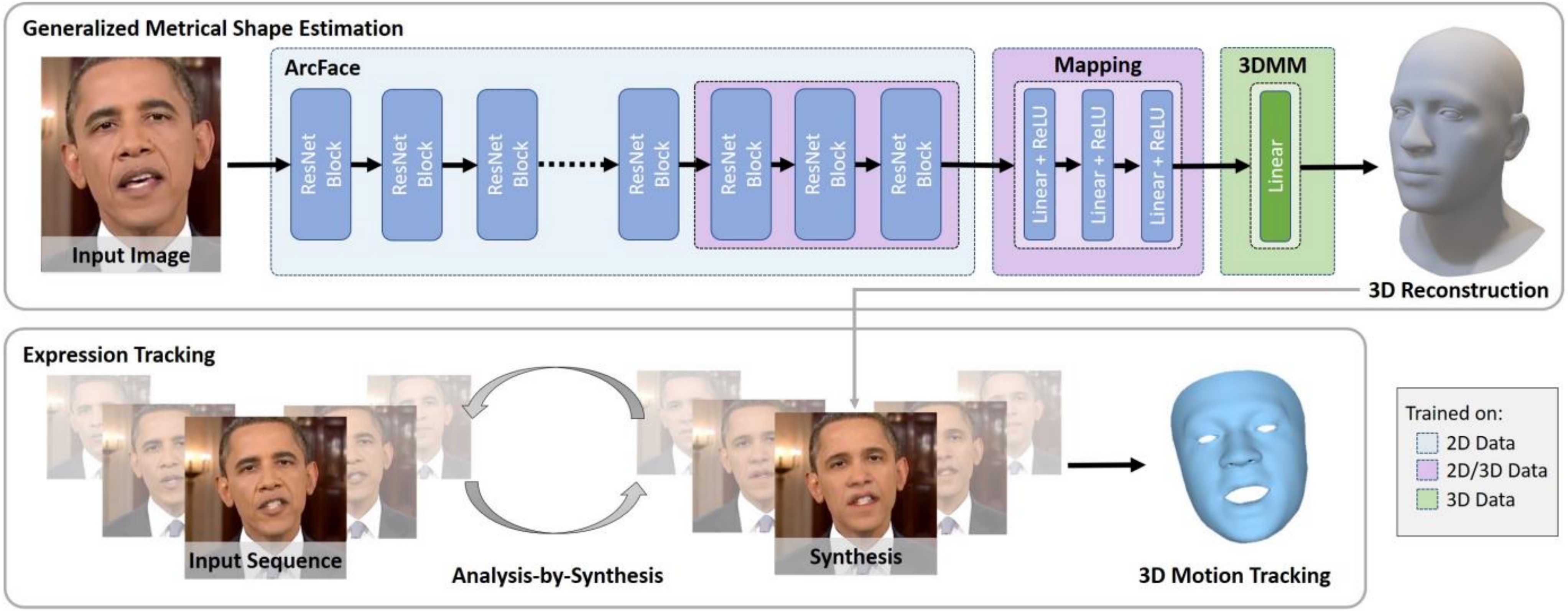


Explicit Morphable Models: FOCUS



Li *et al.* 2023 (To fit or not to fit: Model-based face reconstruction and occlusion segmentation from weak supervision)

Explicit Morphable Models: MICA



Zielonka *et al.* 2022 (MICA: Towards Metrical Reconstruction of Human Faces)

Explicit Morphable Models: Personalized Model

Monocular Video

Grassal *et al.* 2022 (Neural head avatars from monocular RGB videos)

Implicit Morphable Models: Supervision

Supervised

- Photometric loss

Pros

- Expression disentanglement (Editablity applications)

Cons

- Poor latent space
- Doesn't generalize well

Unsupervised (Adversarial)

- GAN loss

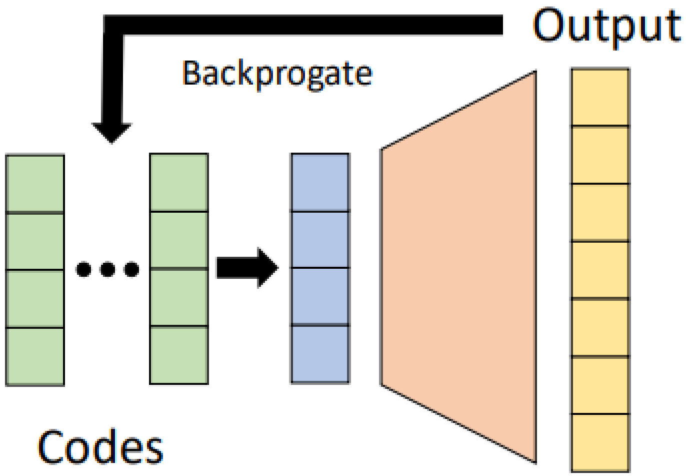
Pros

- Good latent space (good random samples)

Cons

- Accuracy depends on the estimated camera pose distribution

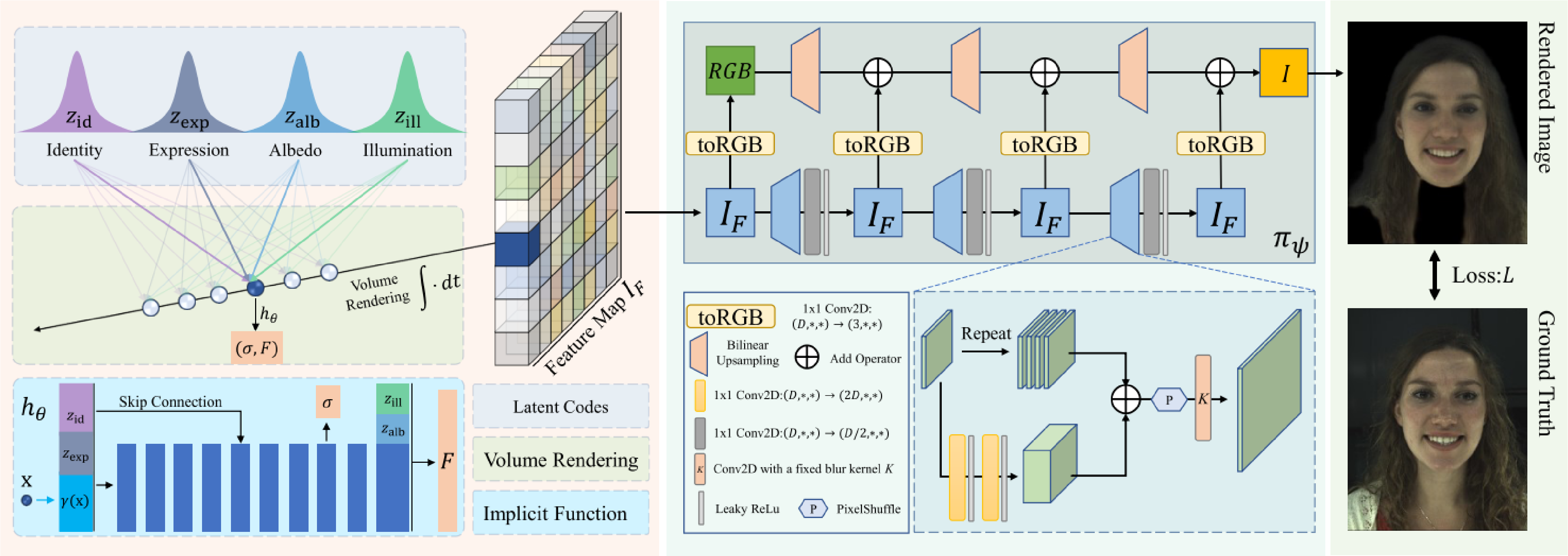
Implicit Morphable Models: Supervised Training



Auto-decoder

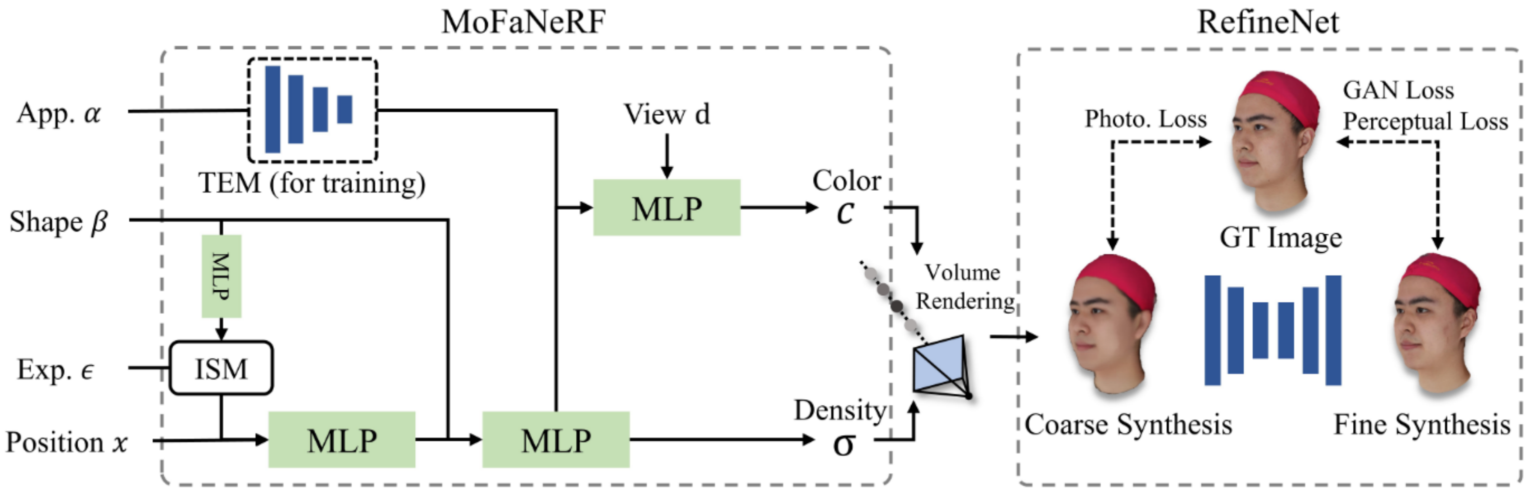
Park *et al.* 2019 (DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation)

Implicit Morphable Models (Supervised Training): HeadNeRF

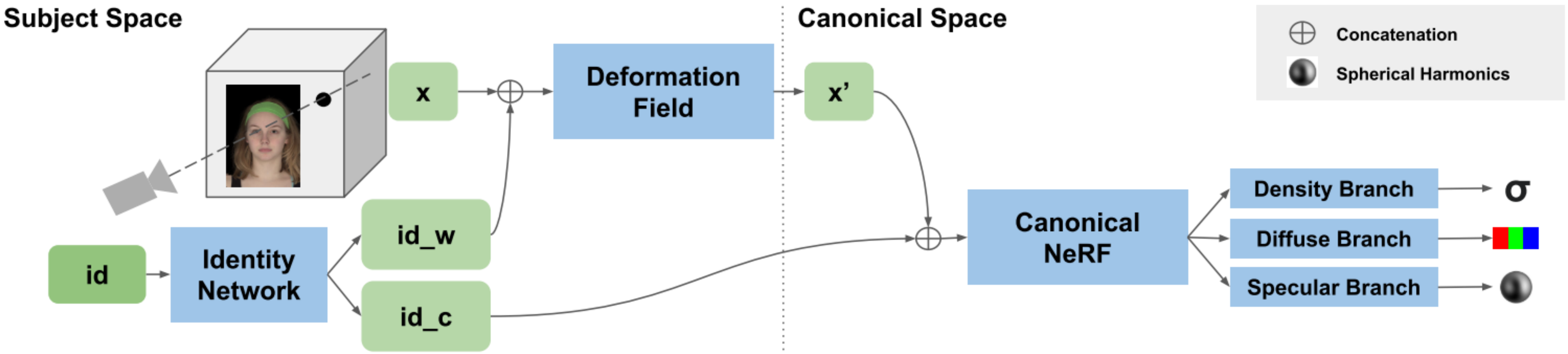


Hong *et al*. 2022 (HeadNeRF: A Real-time NeRF-based Parametric Head Model)

Implicit Morphable Models (Supervised Training): MoFaNeRF, MoRF



Zhuang *et al.* 2022 (MoFaNeRF: Morphable Facial Neural Radiance Field)

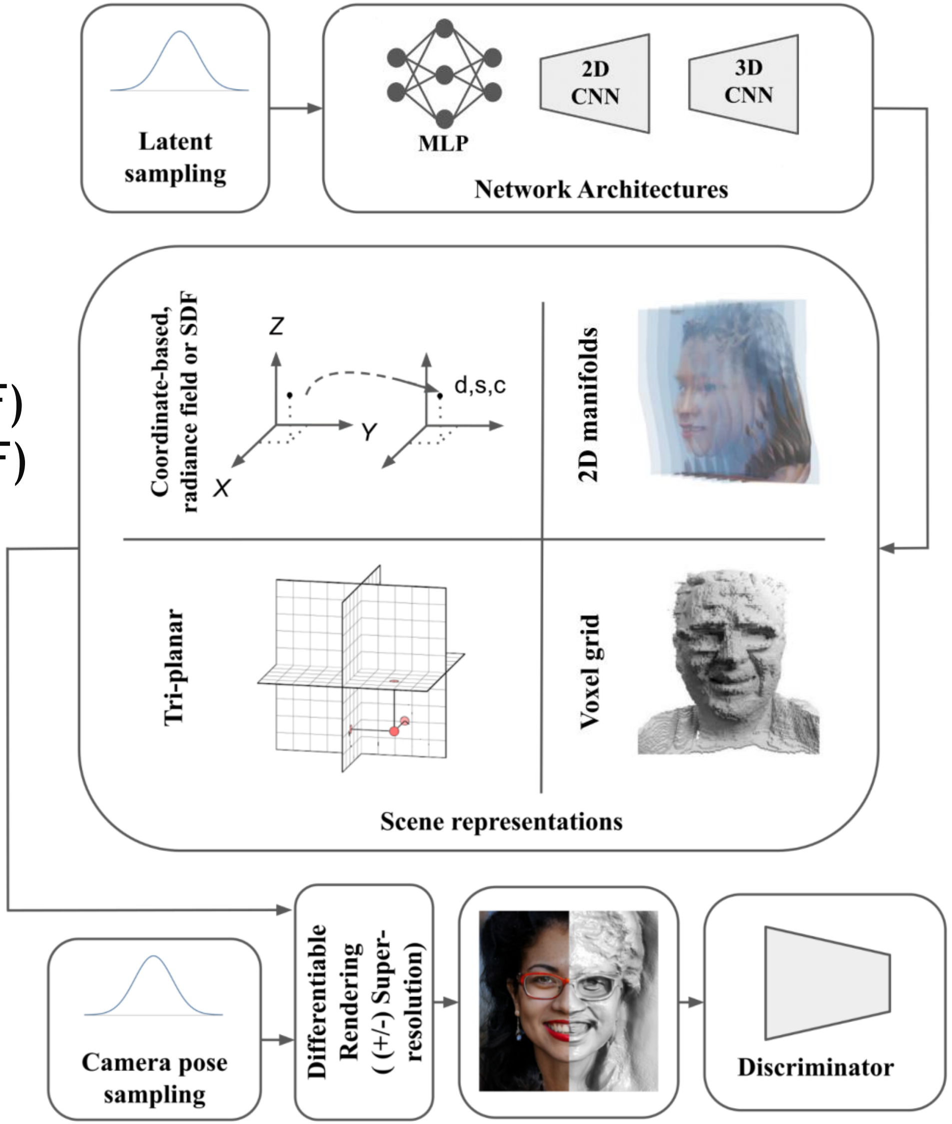


Wang *et al.* 2022 (MoRF: Morphable Radiance Fields for Multiview Neural Head Modeling)

Implicit Morphable Models: Adversarial Training

- Chan *et al.* (pi-GAN)
- Schwarz *et al.* (GRAF)
- Or-El *et al.* (StyleSDF)

- Chan *et al.* (EG3D)

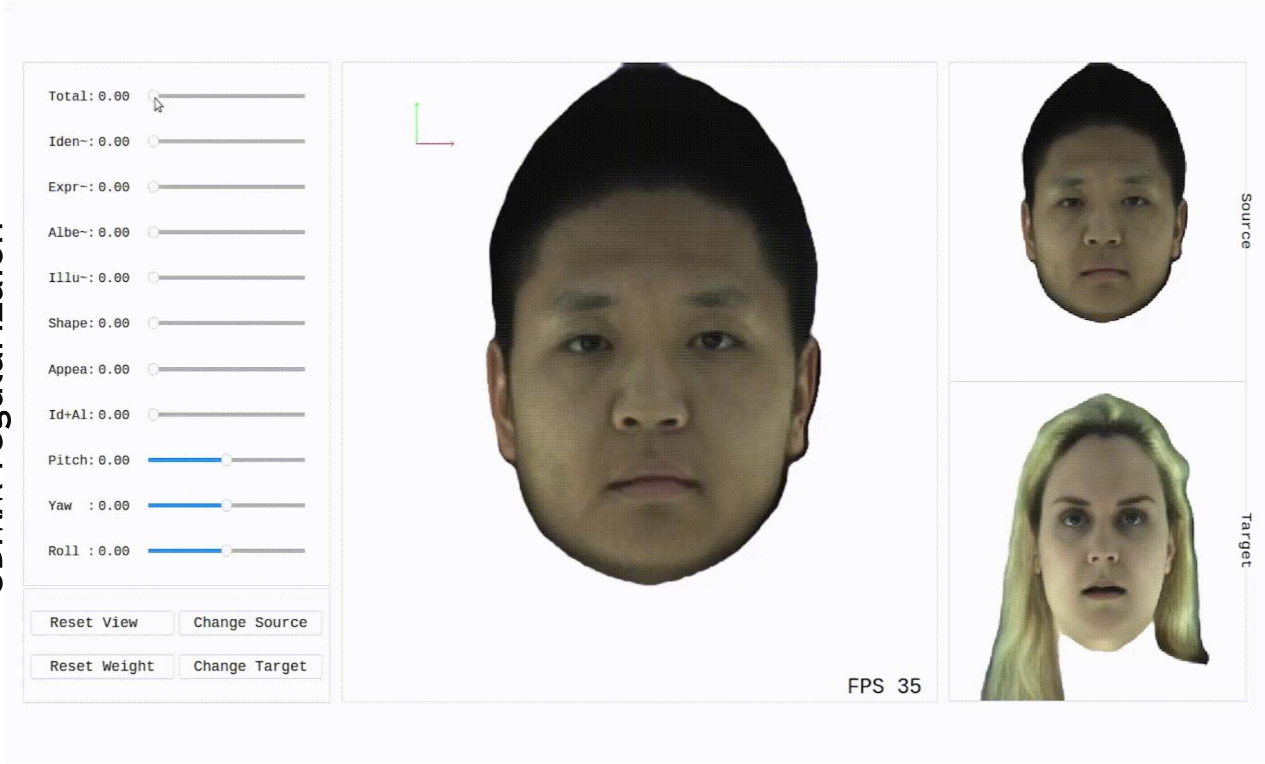


- Deng *et al.* (GRAM)
- Xiang *et al.* (HD-GRAM)

- Schwarz *et al.* (VoxGRAF)

Implicit Morphable Models: Supervised vs. Adversarial

Supervised
(Auto-Decoder) +
3DMM regularizaion



Hong *et al*. 2022 (HeadNeRF)

Sueprvised
(Auto-Decoder)

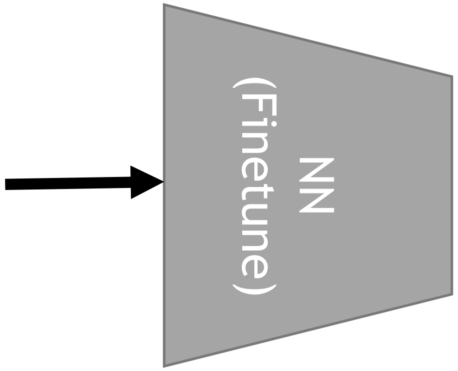
Rebain *et al*. 2022 (LoLNeRF)

Adversarial

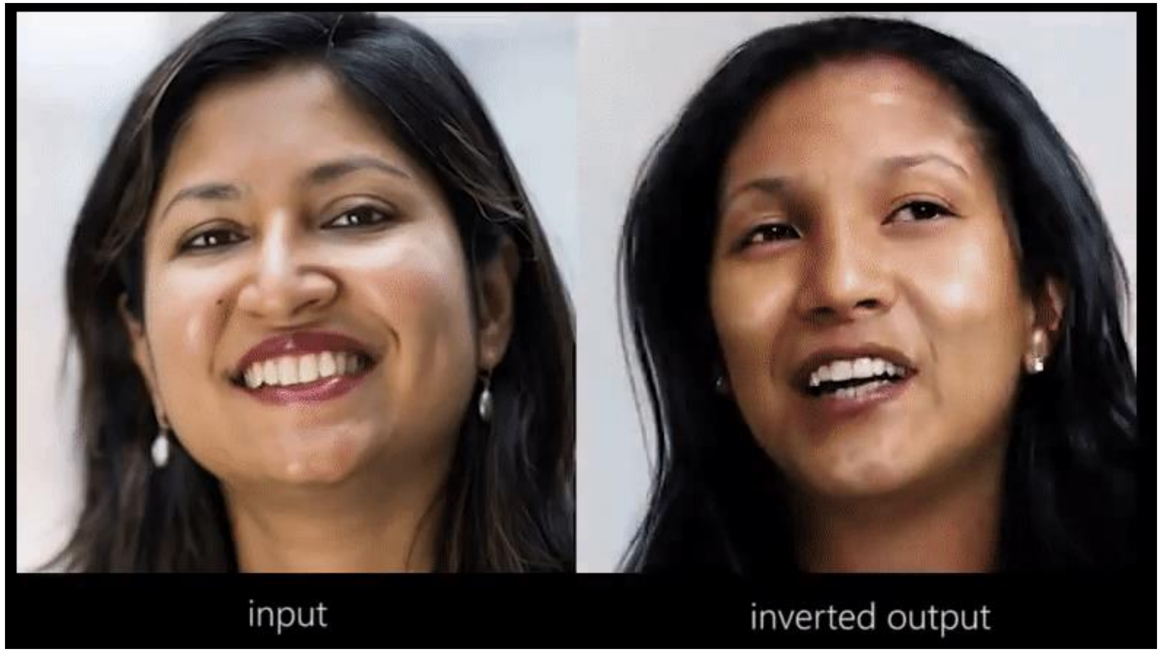
Chan *et al*. 2022 (EG3D)

Implicit Morphable Models: Fitting

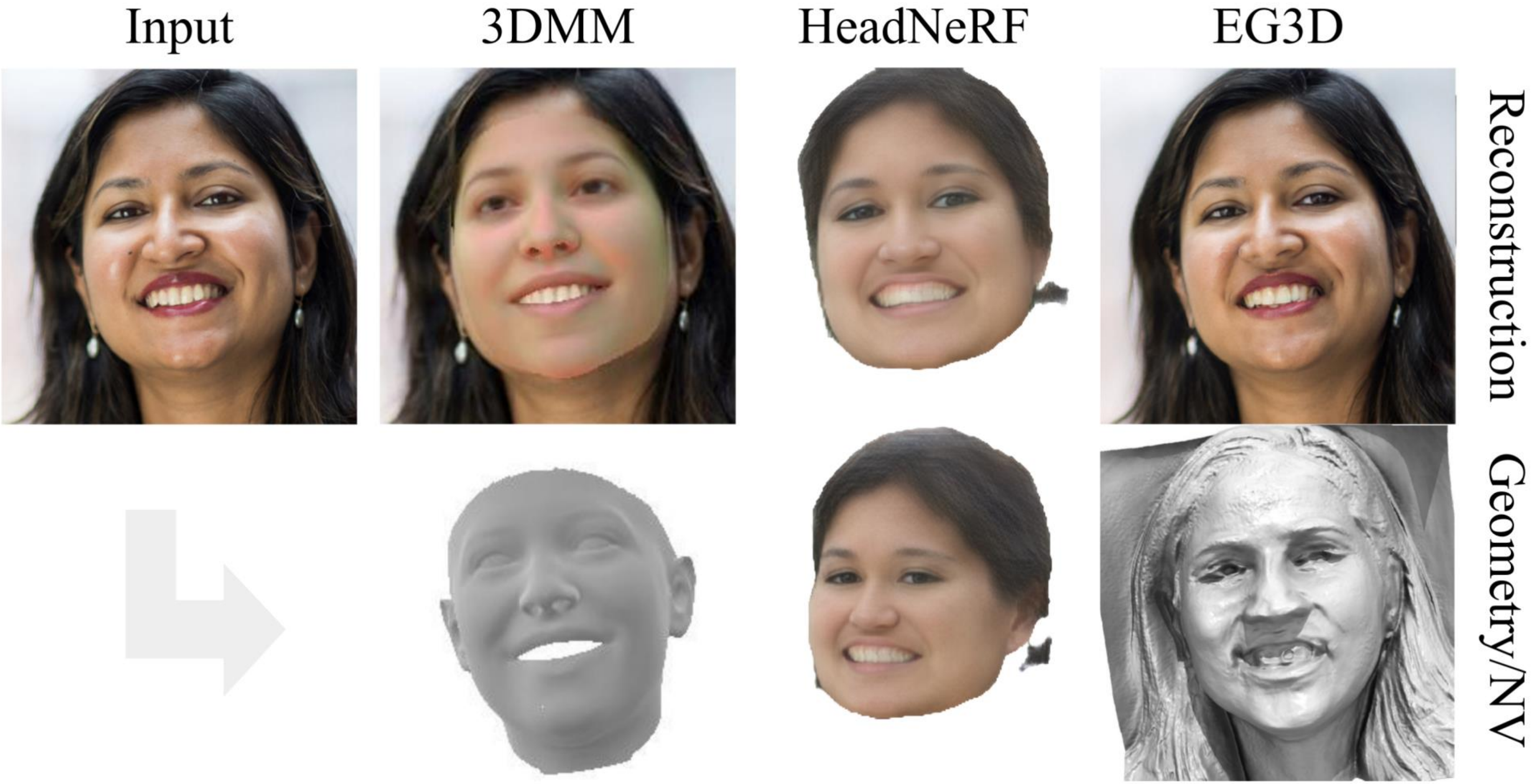
Latent
(Optimize)



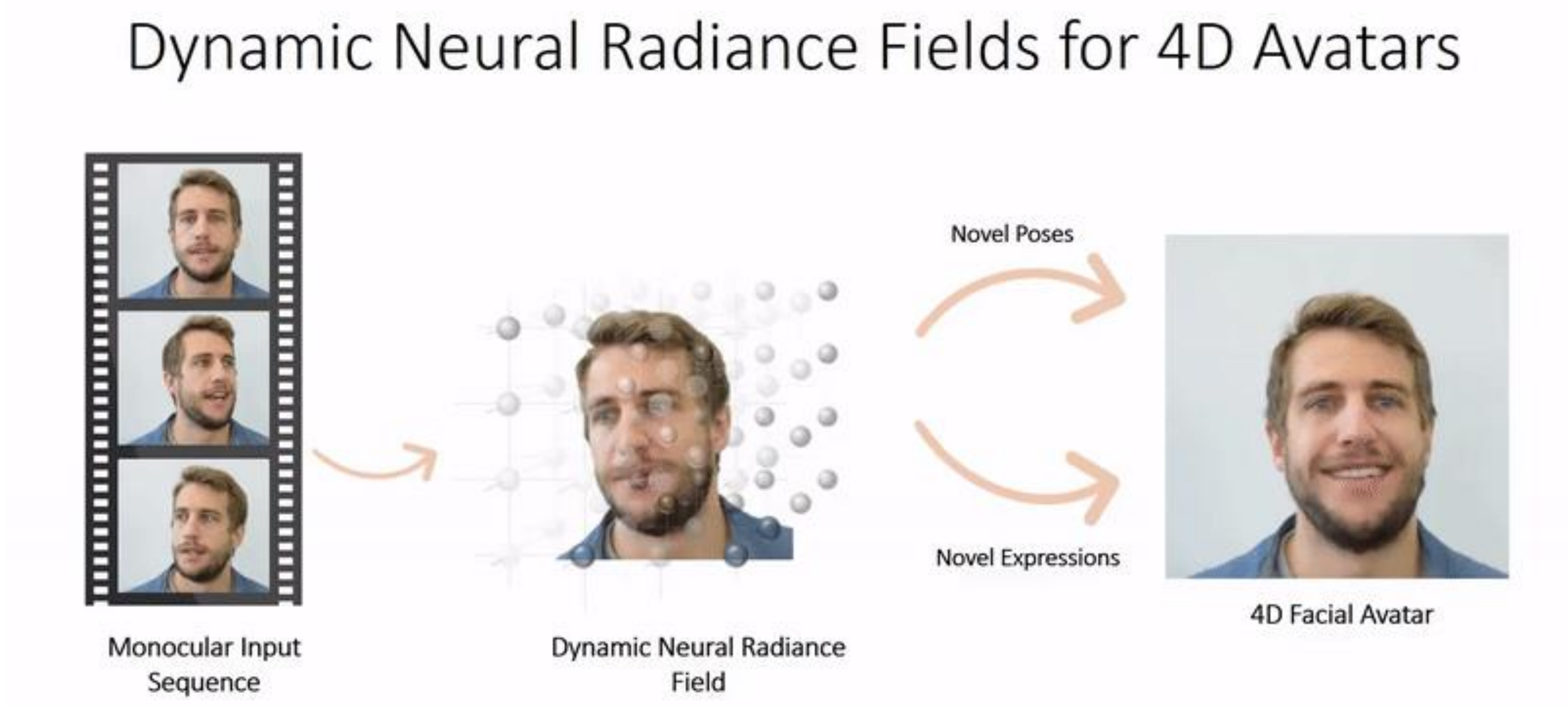
Photometric
Loss



Implicit Morphable Models: Comparison

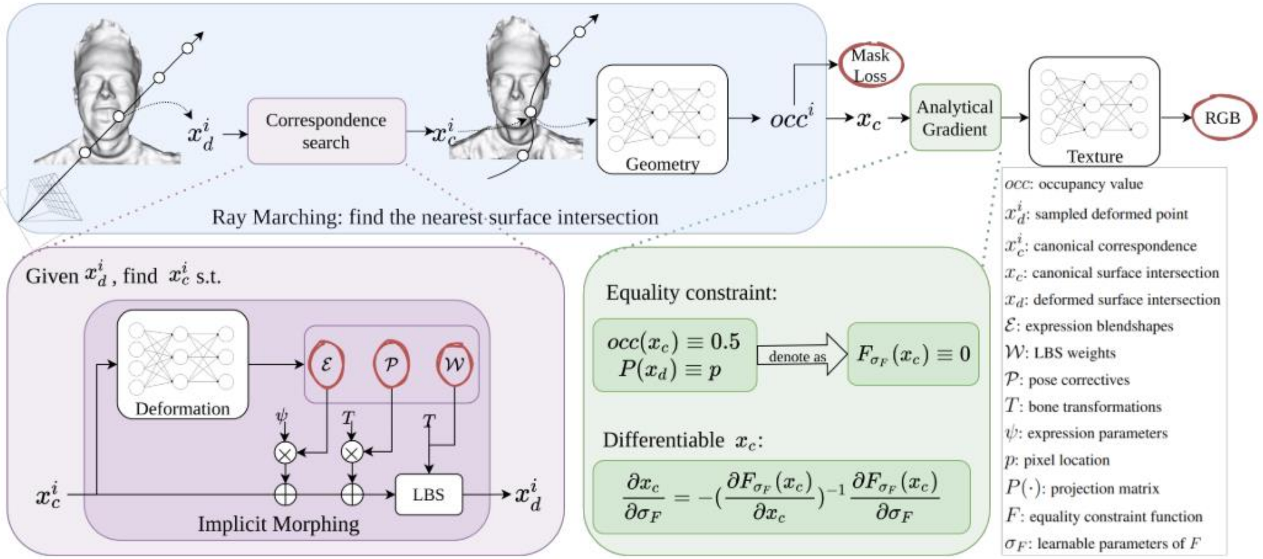


Implicit Morphable Models: Person-Specific Model

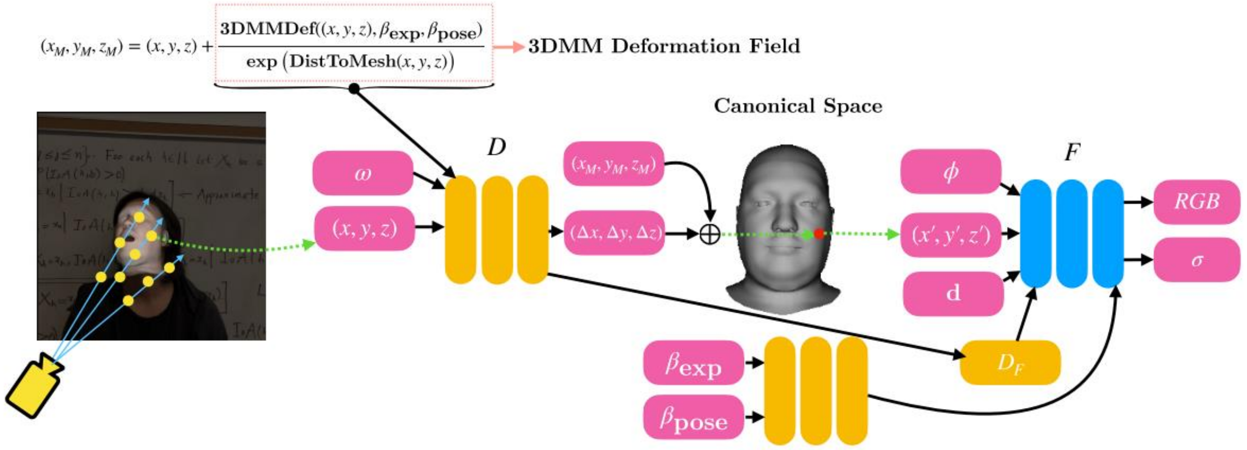


Gafni *et al*. 2021 (Dynamic neural radiance fields for monocular 4D facial avatar reconstruction)

Implicit Morphable Models: Person-Specific Model



Zheng *et al.* 2022 (I M Avatar: Implicit morphable head avatars from videos)



Athar *et al.* 2022 (RigNeRF: Fully controllable neural 3d portraits)

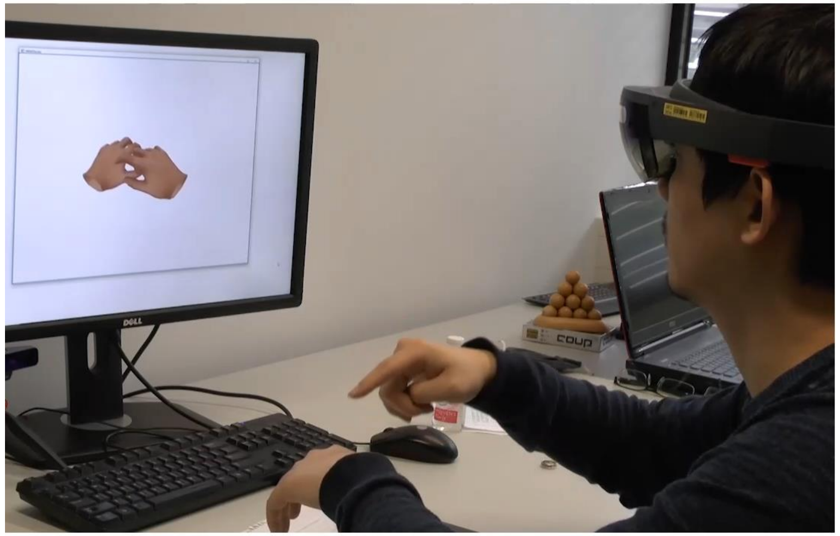
Conclusion

- Explicit models struggle with complex topology and finer details
- Recent implicit models that use neural networks to build prior are over-parameterized
- Metrically accurate generative models
- No non-person specific implicit model based methods exist that take advantage of video dataset

3.4 Hands

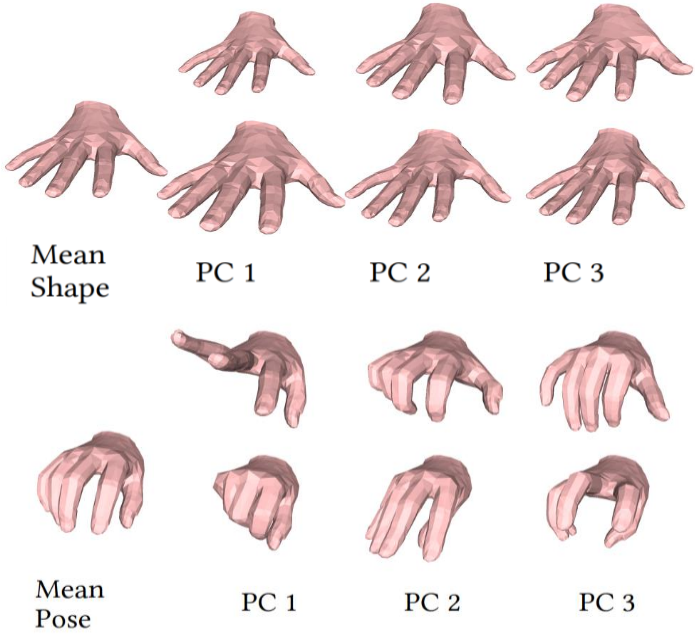
Hands

- Highly articulated: Pose-dependent deformations
- Severe self-occlusions
- Shape from Template?
  - Requires 3D template
  - Not robust to occlusions
- Parametric 3D hand model as a prior, *e.g.* Romero *et al.* 2017 (MANO)

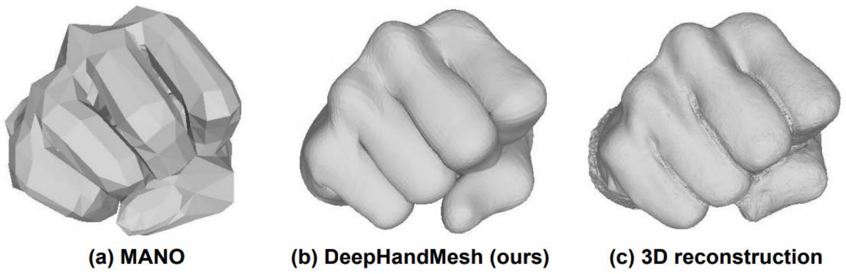


VR View

3D Hand Models



Statistical hand model: Romero *et al.* 2017 (MANO)



Personalized high-res model:
Moon *et al.* 2020 (DeepHandMesh)

Input image Re-projected hand mesh 3D (novel view)

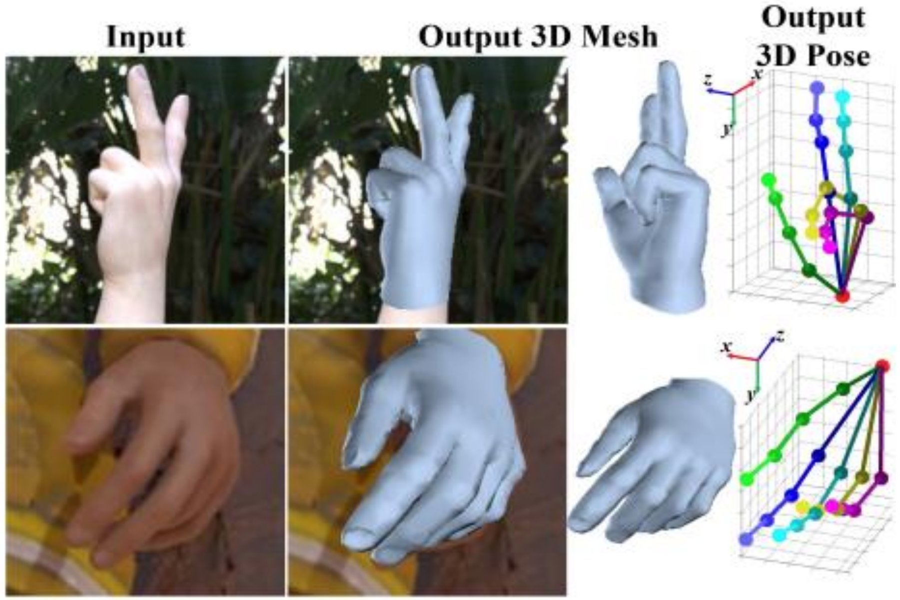
Hand texture model: Qian *et al.* 2020 (HTML)

Input image Re-projected hand mesh 3D (novel view)

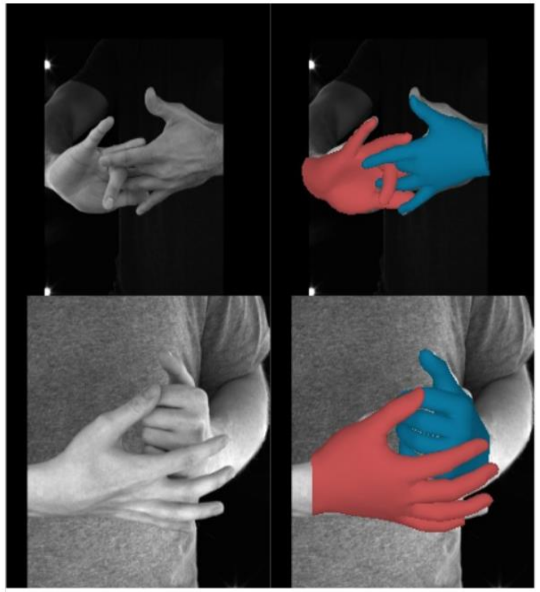
Implicit hand model: Corona *et al.* 2022 (LISA)

Hands

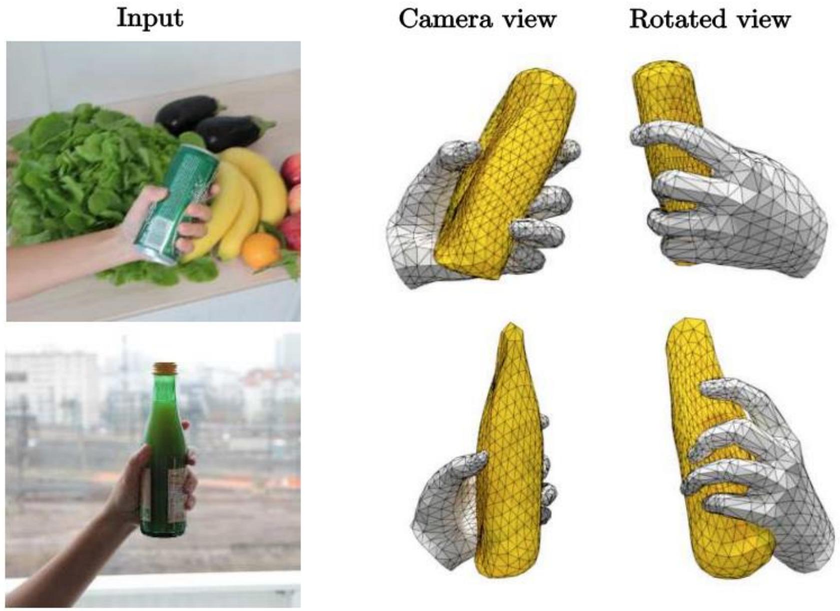
Different scenarios:



Single hand

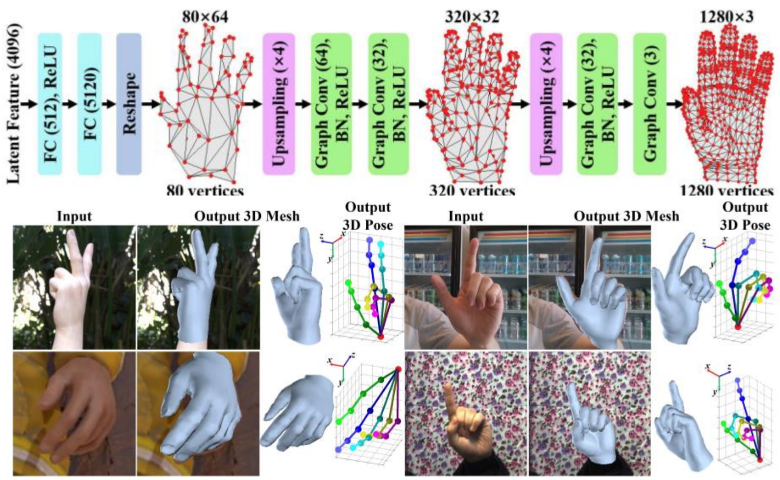


Two interacting hands

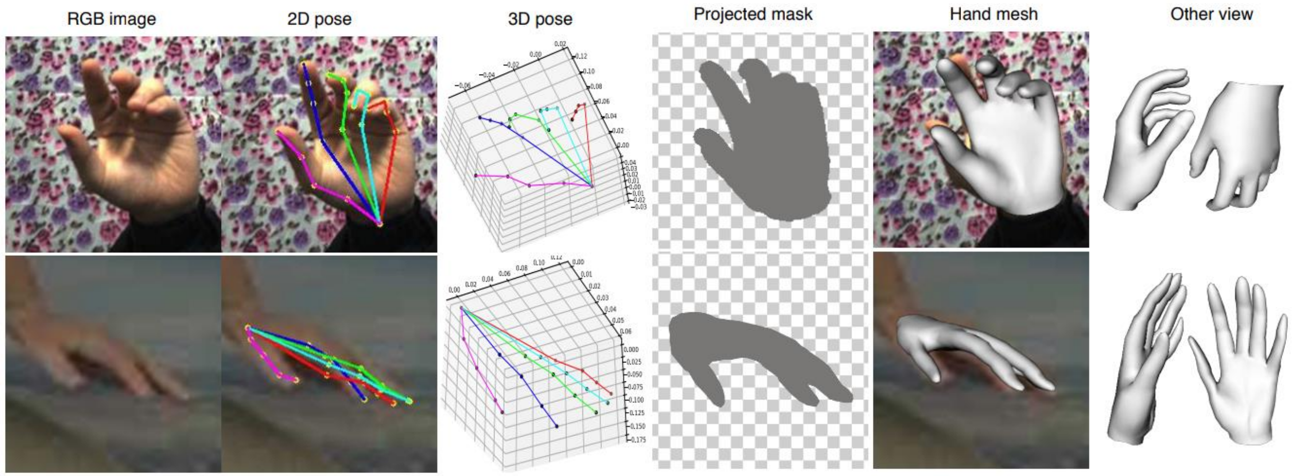


Hands and an object

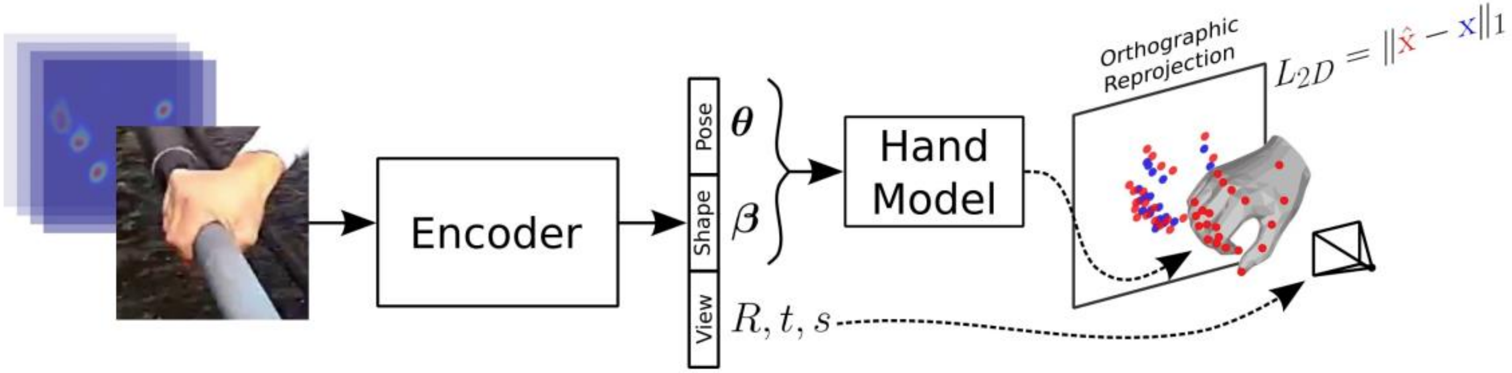
Single Hands: Regression of MANO Parameters



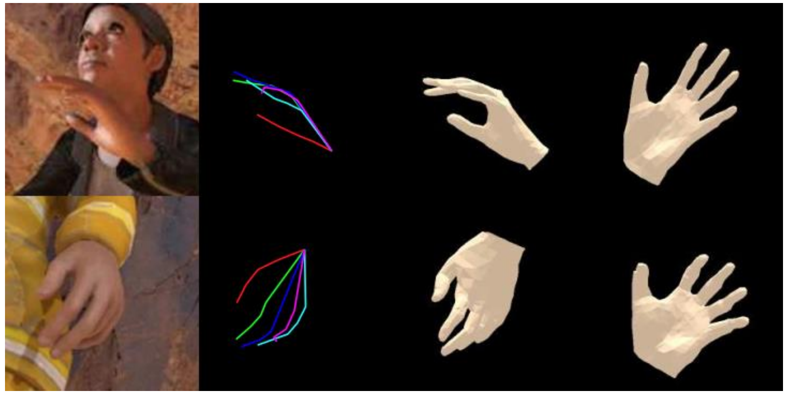
Ge *et al.* 2019



Zhang *et al.* 2019

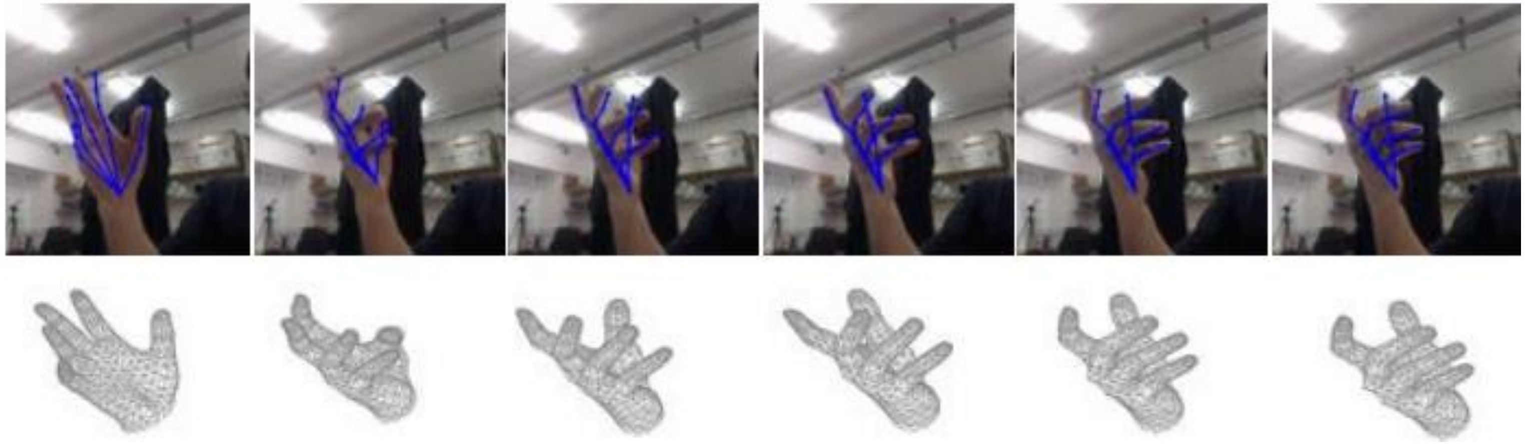


Boukhayma *et al.* 2019



Boukhayma *et al.* 2019

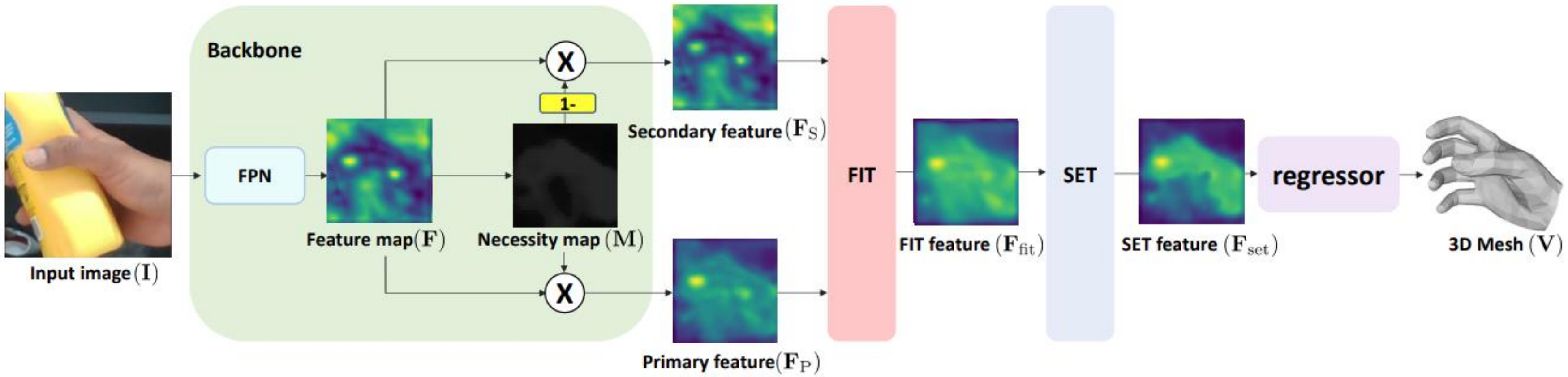
Single Hands



Yang *et al.* 2020 (SeqHAND)

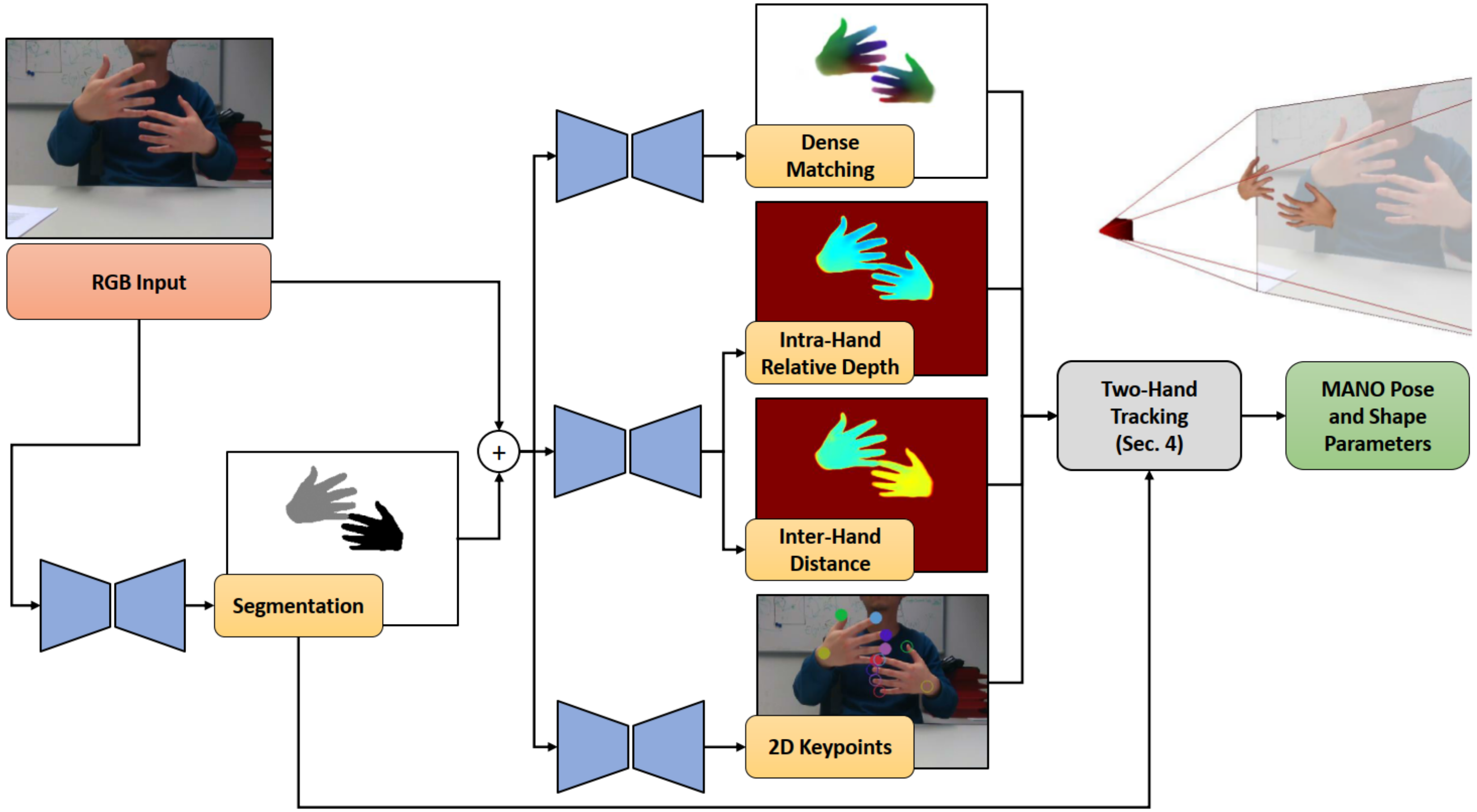
*Learning framework with a **temporal** component*

*Hand mesh estimation network **robust to occlusions***

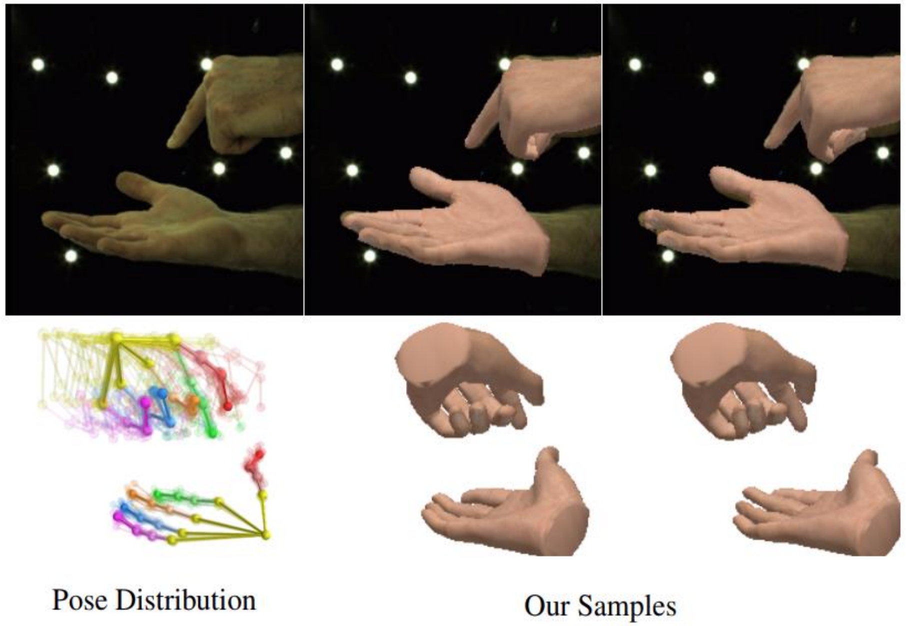


Park *et al.* 2022 (HandOccNet)

Two Interacting Hands

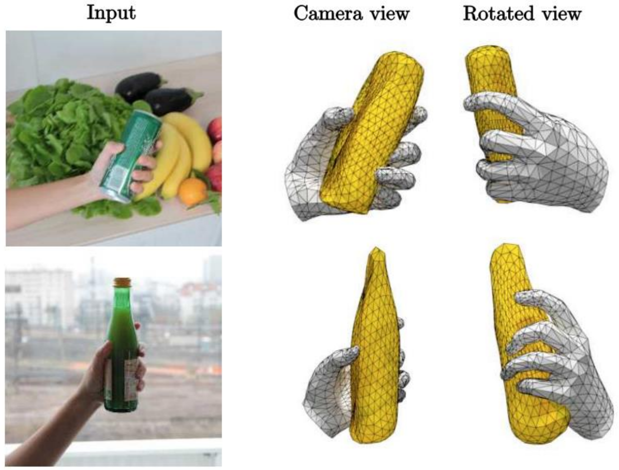


Wang *et al*. 2020 (RGB2Hands)

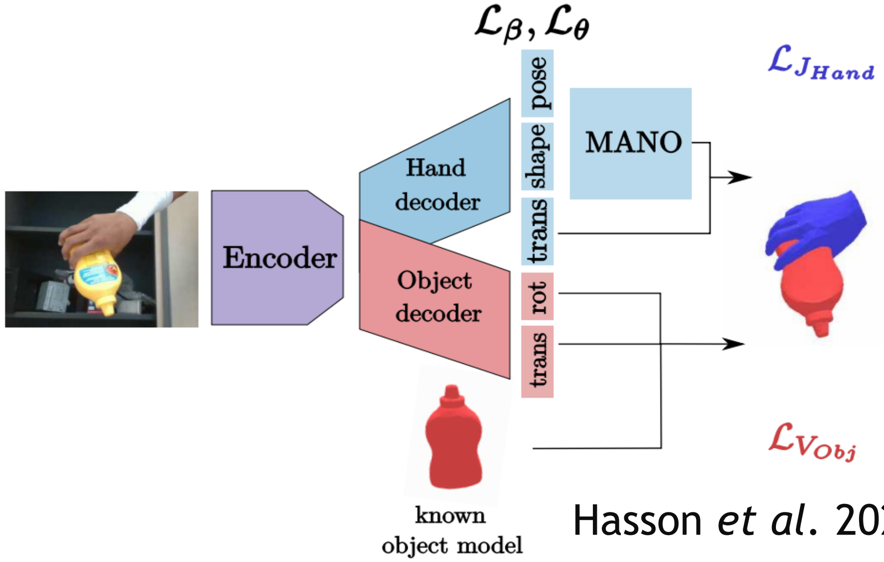


Wang *et al*. 2022 (HandFlow)

Hands and Objects

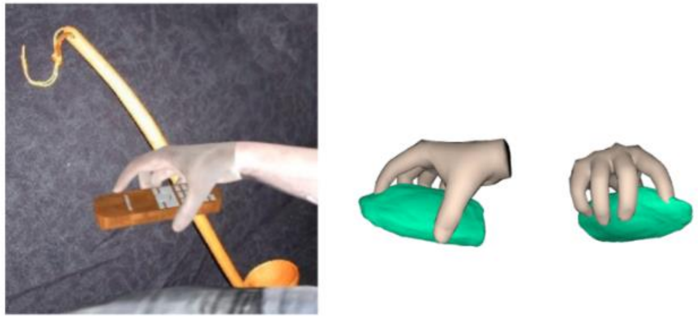


Hasson *et al*. 2019



Hasson *et al*. 2020

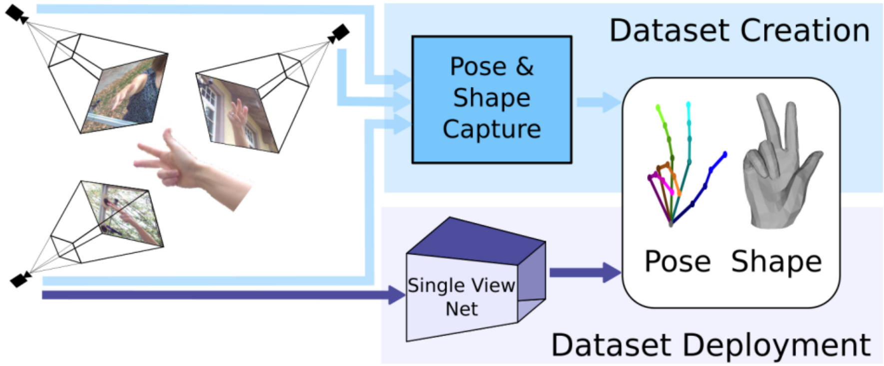
- Joint reconstruction
- Known 3D object model
- Local object shape drives hand articulations



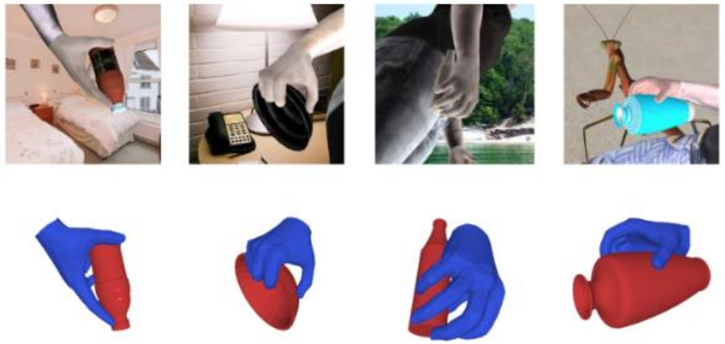
Karunratanakul *et al*. 2020

Ye *et al*. 2022

Datasets for 3D Hand Pose Estimation



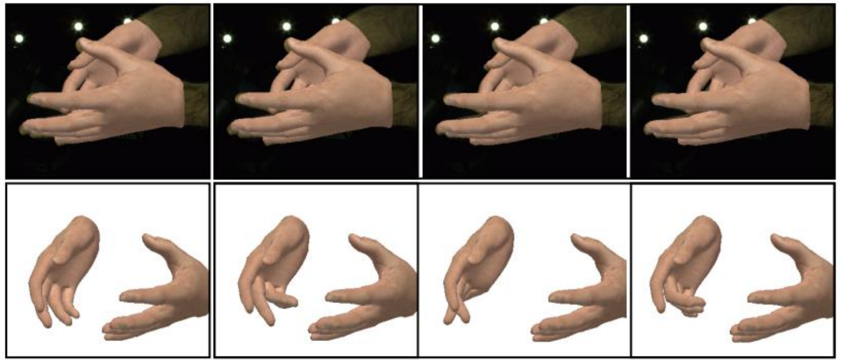
Zimmermann *et al*. 2019 (FreiHAND)



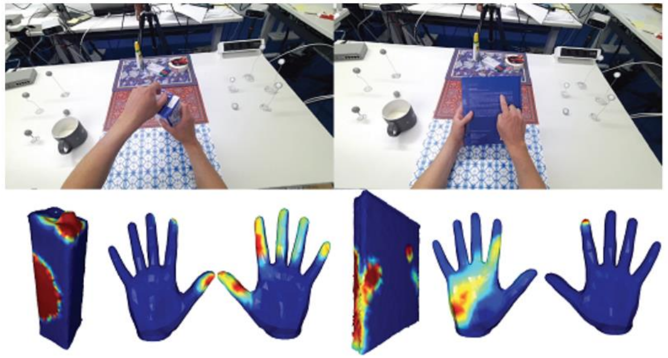
Hasson *et al*. 2019 (ObMan)

Chao *et al*. 2021 (DexYCB)

Moon *et al*. 2020 (InterHand2.6M)



Wang *et al*. 2022 (MultiHands)



Kwon *et al*. 2021 (H2O)

Future Directions

- More geometric and pose-dependent details
  - Nails, hair and blood vessels
- Hands + *deformable* objects
- Relighting of hands under various illuminations
- Improved mesh collisions

3.5 Animals

Animals

- Task-specific motivation: Behavior analysis
- Here: Parametric animal models, not just a template
- Small area (about ten papers) – Why?
  - No good datasets: Capturing animals is more difficult than capturing humans (lack of control, much wider variety)
  - But: SMAL parametric model (Zuffi *et al.* 2017) from quadruped toy animals
- Variety of works:
  - Going beyond SMAL shape space (Zuffi *et al.* 2018, Li *et al.* 2021)
  - Video input (Biggs *et al.* 2018)
  - Train on synthetic data, test on real data (Zuffi et al. 2019)
  - Building SMAL-style models from 2D inputs
    - Dogs (Biggs *et al.* 2020), "breed-aware" (Rüegg *et al.* 2022)
    - Birds: Single species (Badger *et al.* 2020), multiple species (Wang *et al.* 2021)
  - Retrieve good bird template, then deform (Wu *et al.* 2022)

Wang *et al.* 2021

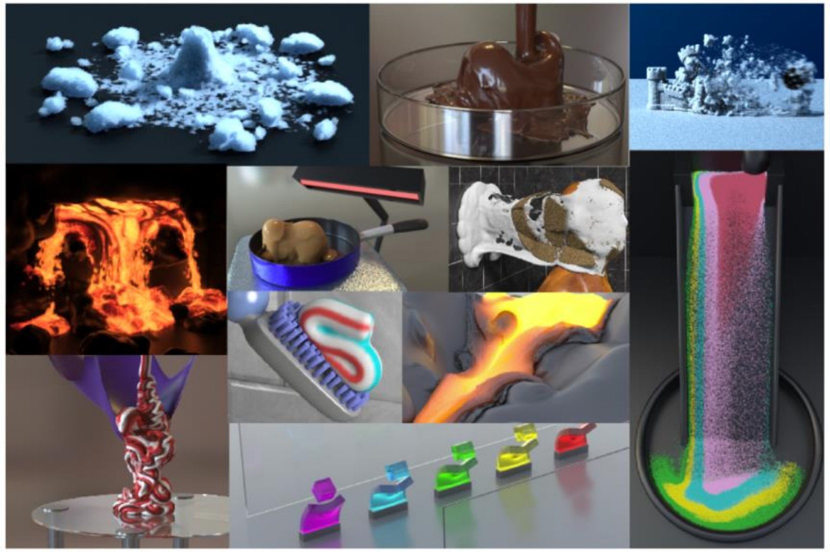
4

Emerging Areas

4.1 Physics

Physics-Based Reconstruction

Physics-based simulation of deformable objects



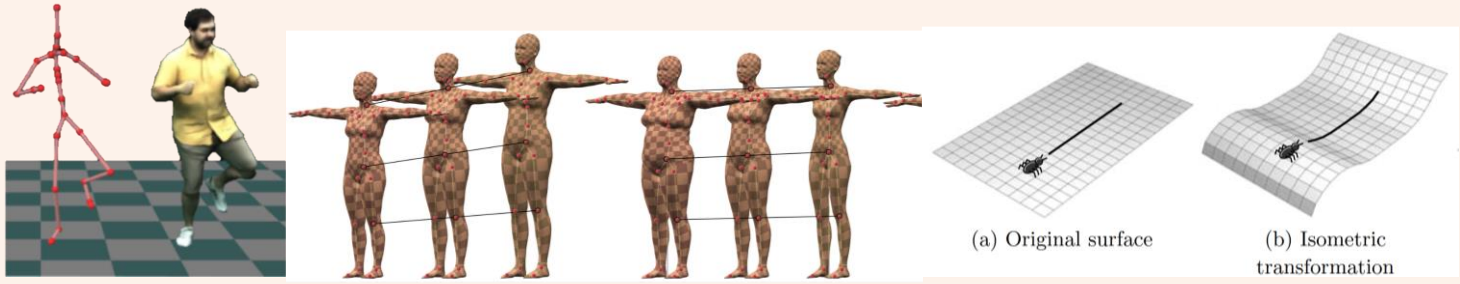
Jiang *et al.* 2016

Render

Reconstruction

Physically-based rendering

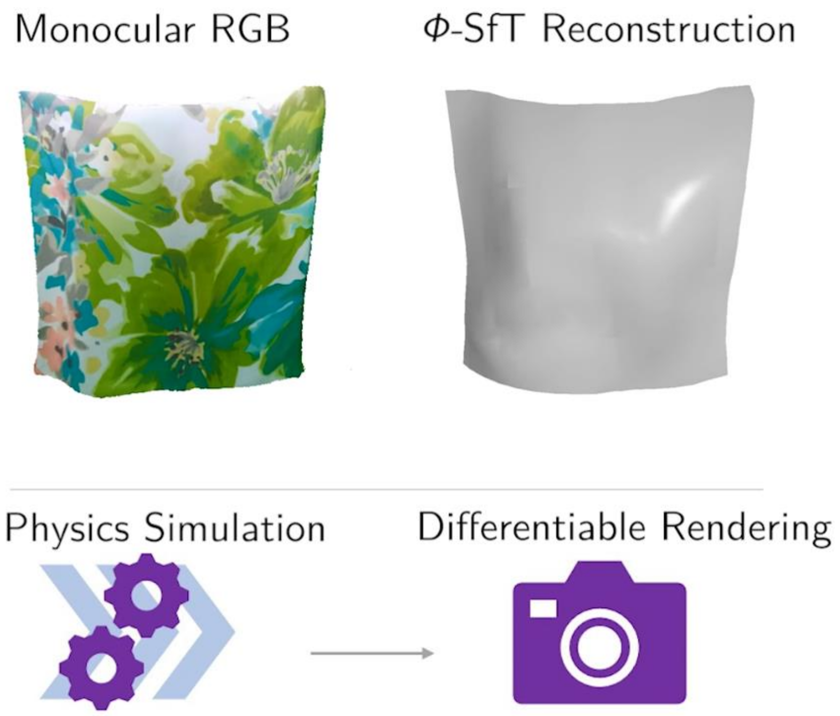
Phar *et al.* 2023(PBRT)



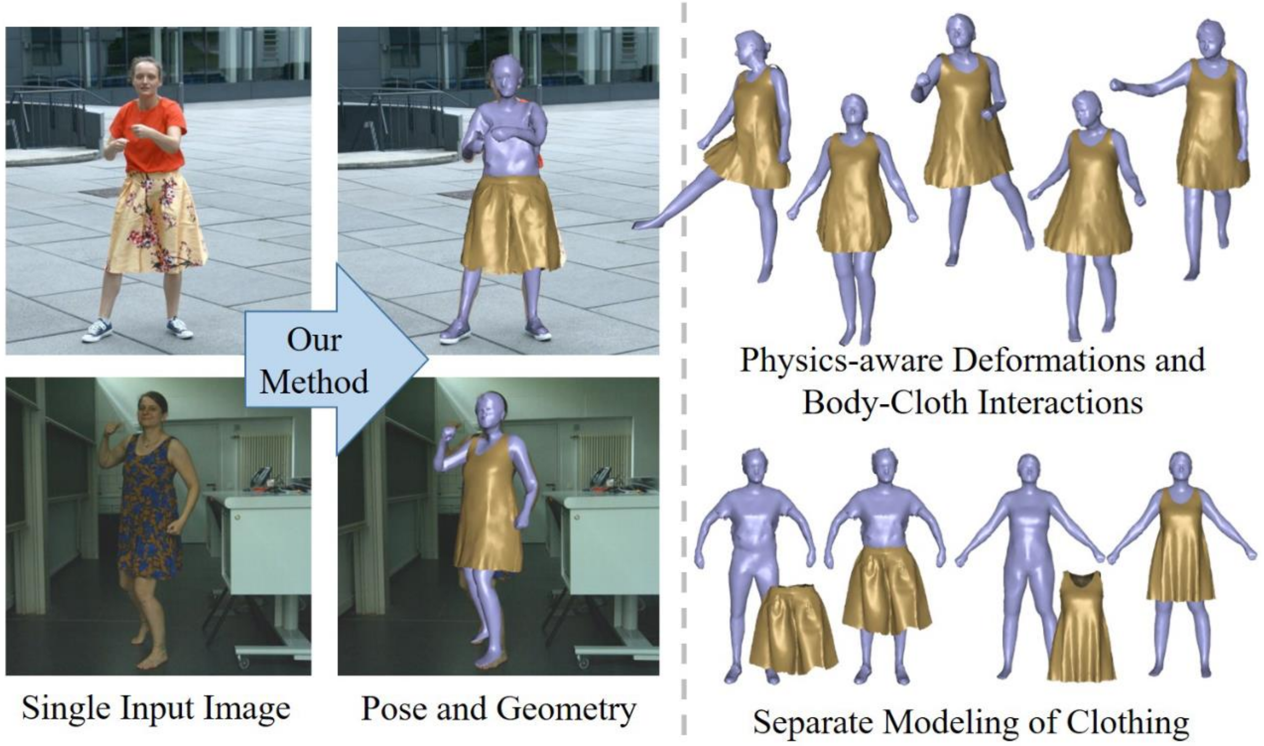
Geometric approximation of physical behavior

Physics-Based Reconstruction

Physics simulation as **soft constraint**



Kairanda *et al.* 2022 (φ-SfT)

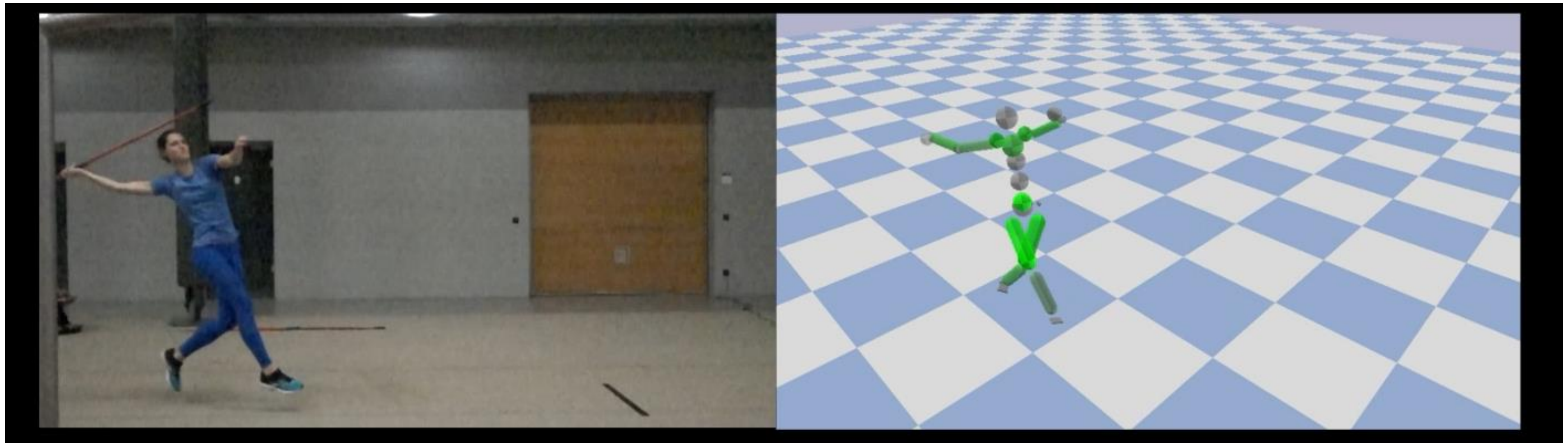


Li *et al.* 2021

Physics simulation as **hard constraint**

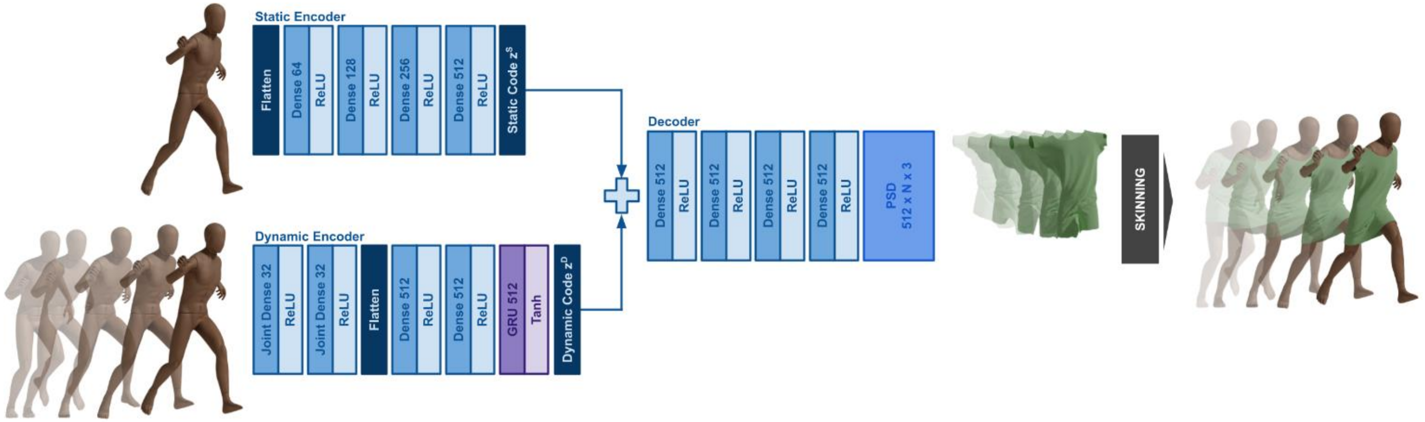
Physics-Based Reconstruction

- Last decade: Learning-based methods
- Emerging trend: Physics + learning
  - *E.g.* sparse reconstruction – human motion capture



Shimada *et al.* 2021 (Neural PhysCap)

Extension to dense?



Bertiche *et al.* 2022 (Neural Cloth Simulation)

Similar ideas for reconstruction?

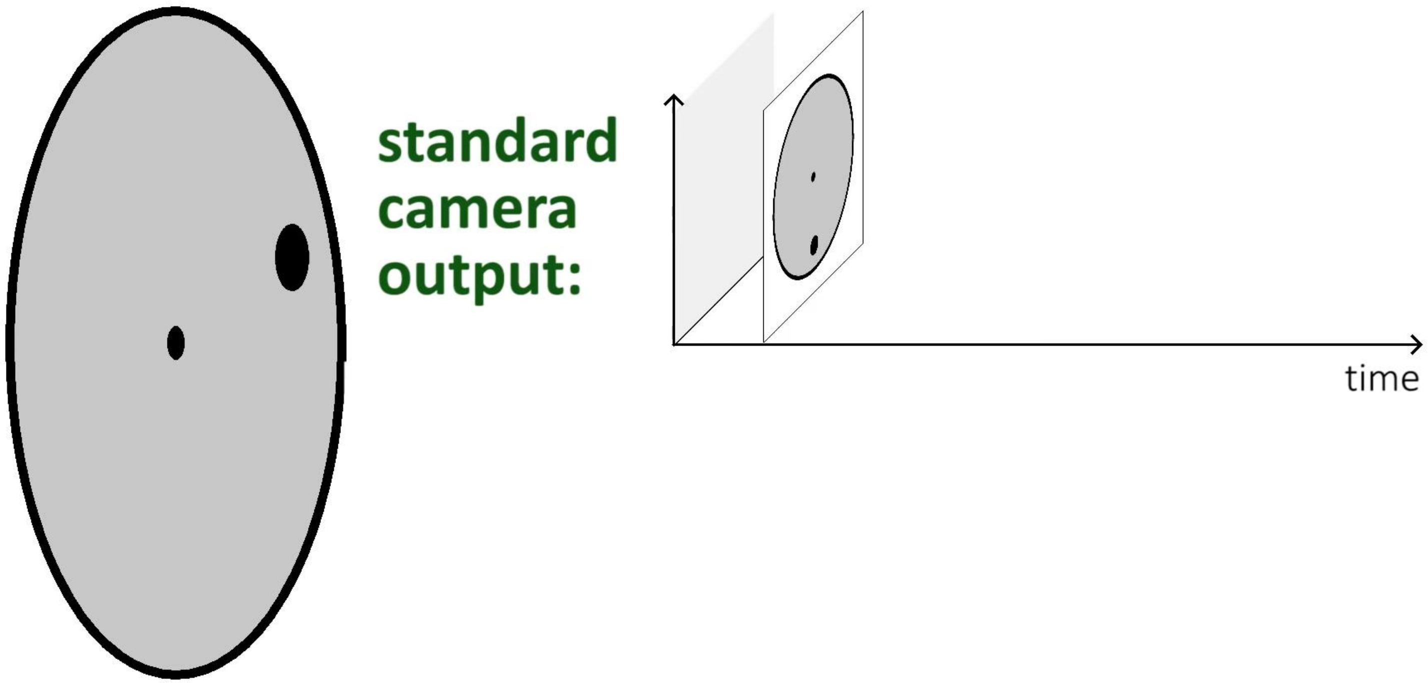
Physics-Based Reconstruction: Future Directions

- Full physics modelling of complex objects
  - *E.g.* human skin, muscles, hair and clothing
- Need to account for many physical phenomena
  - *E.g.* contacts, collisions, elasticity, plasticity or fractures
- Integration with neural methods
  - Fast inference, memory efficient
  - Physics as loss functions (Raissi *et al*. 2019)
  - Differentiable physics simulation as a layer (Liang *et al*. 2019)

4.2
Event Cameras

Event Cameras

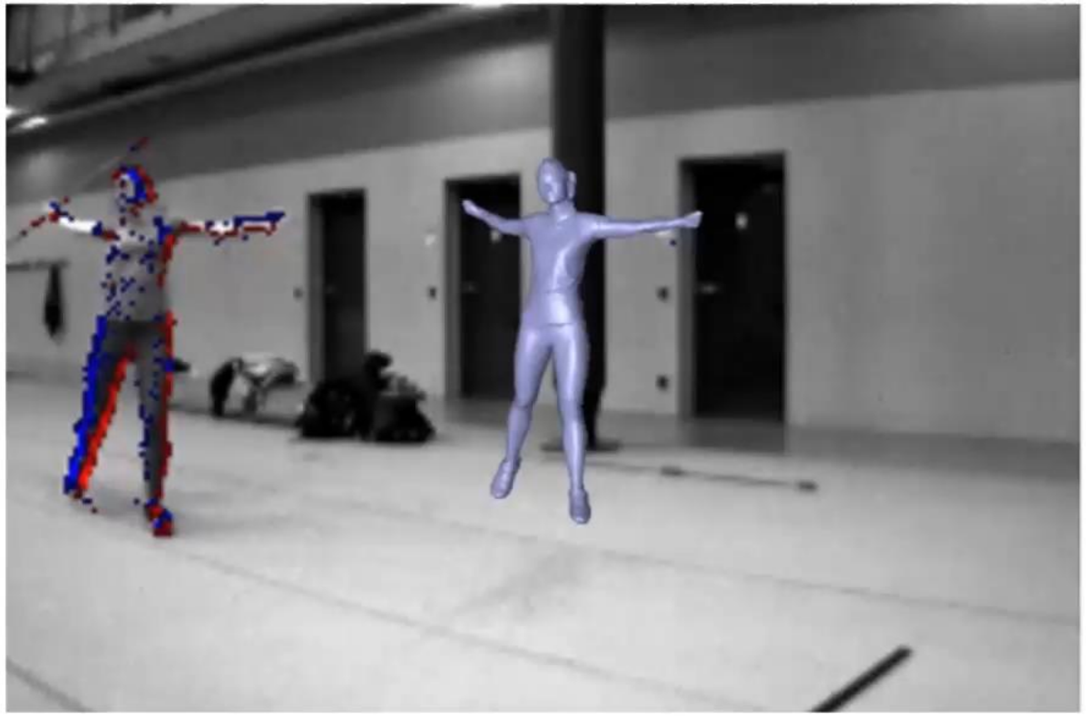
Events: Changes in brightness, recorded asynchronously per-pixel



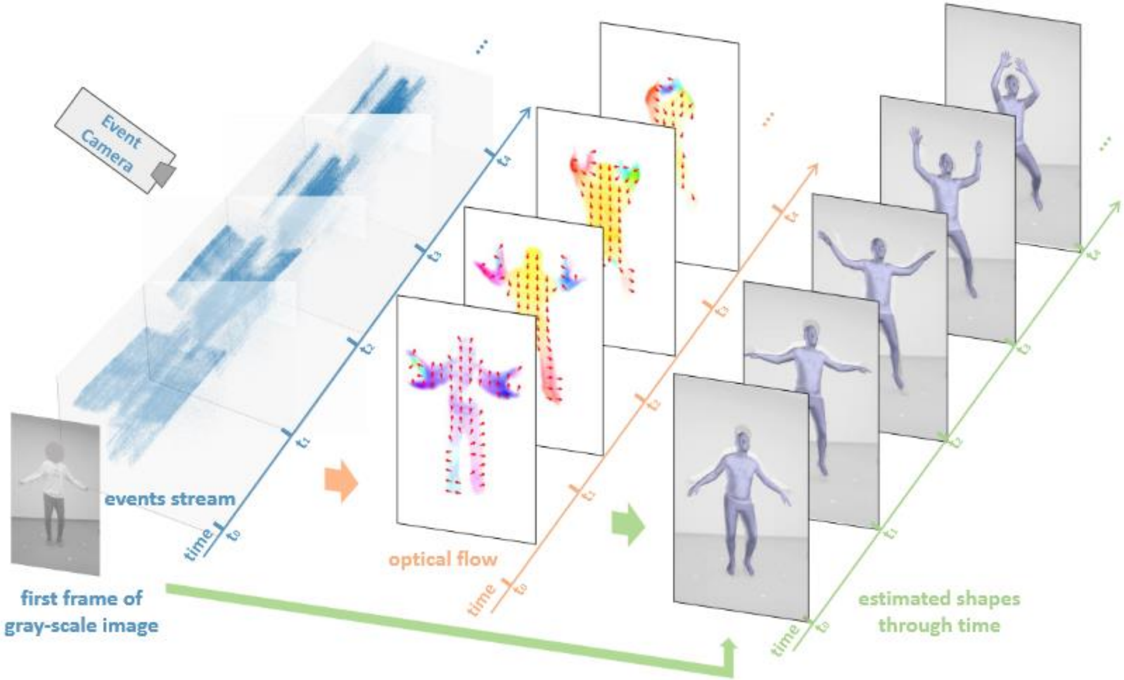
- No motion blur
- HDR

Source: https://youtu.be/LauQ6LWTkxM?t=30

Reconstruction with Event Cameras: State of the Art

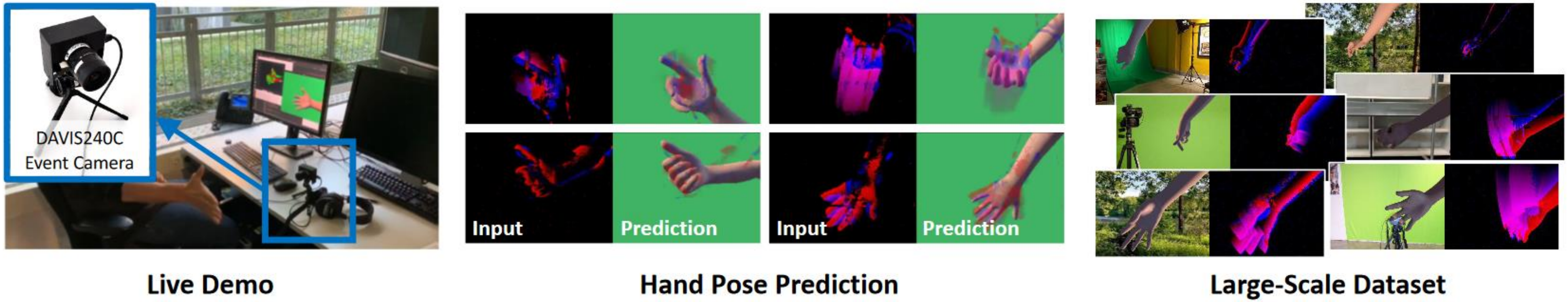


Xu *et al*. 2020 (EventCap)



Zou *et al*. 2021 (EventHPE)

Reconstruction with Event Cameras: State of the Art



Rudnev *et al*. 2021 (EventHands)

Comparison to reconstruction with RGB:

+ Better synthetic-to-real generalization
+ High-speed motion reconstruction using much lower bandwidth
- Single or few events are not sufficient for reconstruction

5

Open Challenges

Open Challenges

Large Scale

- Some NeRF-based methods handle nearby static background

Multiple Objects

- Explicit handling of multiple objects is in its infancy (Menapace *et al*. 2022)

Data Bias

- Datasets do not reflect real-world appearance distribution of people → Indirect bias in method design via evaluation, direct bias in learning-based methods
- Benchmarks can quantify bias (Feng *et al*. 2022)

Model Variety

- Morphable and parametric models assume able-bodied individuals
- Also do not account for individualistic appearance variation like tattoos

Editability

- Deformations make editing hard, especially fine details like wrinkles
- Comparatively easy for meshes
- Very difficult with neural scene representations

Real-Time Performance

- Some category-specific methods are real time (Tewari *et al*. 2018)
- Related settings like RGB-D or static RGB reconstruction are real time

6

Social Implications

Social Implications

- Many upsides as discussed previously, but some potential social downsides

Environment	Privacy and Consent	Inclusiveness	Authoritativeness	Accessibility
• GPUs need energy, special materials and production	• Need to be considered for human data • Editability can lead to malevolently modified content → Countermeasures are an active research area	• Need to cover a wider range of variation among people (see Open Challenges)	• Specialized methods (*e.g.* for faces) can be reliable • In legal contexts, general methods are unreliable for occluded regions	• Papers, code, datasets, RGB cameras easily obtainable • GPU resources somewhat accessible via cloud services

7

Conclusion

Conclusion

- Largest impact via neural networks:
  Deep learning, differentiable and neural rendering
  → New fields and problem settings now tractable
- Current state:
  - General methods: Still in early phase but going beyond SfT and NRSfM seems promising
  - Humans and faces: Maturing, photo-realism in most settings within reach
  - Hands and animals: Still early, lots of problems remain unaddressed
- Lots of possibilities for the future:
  - Better data via larger, more diverse datasets?
  - Better geometry and appearance via neural scene representations?
  - Better deformations via physics?
  - Better robustness via event cameras?
  - Continued shift from appearance towards learned features via vision transformers, *e.g.* Oquab *et al.* 2023 (DINOv2)?
  - Completely new trends like diffusion, *e.g.* Jakab *et al.* 2023 (Farm3D)?

Thank You!

State of the Art in Dense Monocular Non-Rigid 3D Reconstruction

Edith Tretschk* Navami Kairanda* Mallikarjun B R Rishabh Dabral Adam Kortylewski
Bernhard Egger Marc Habermann Pascal Fua Christian Theobalt Vladislav Golyanik