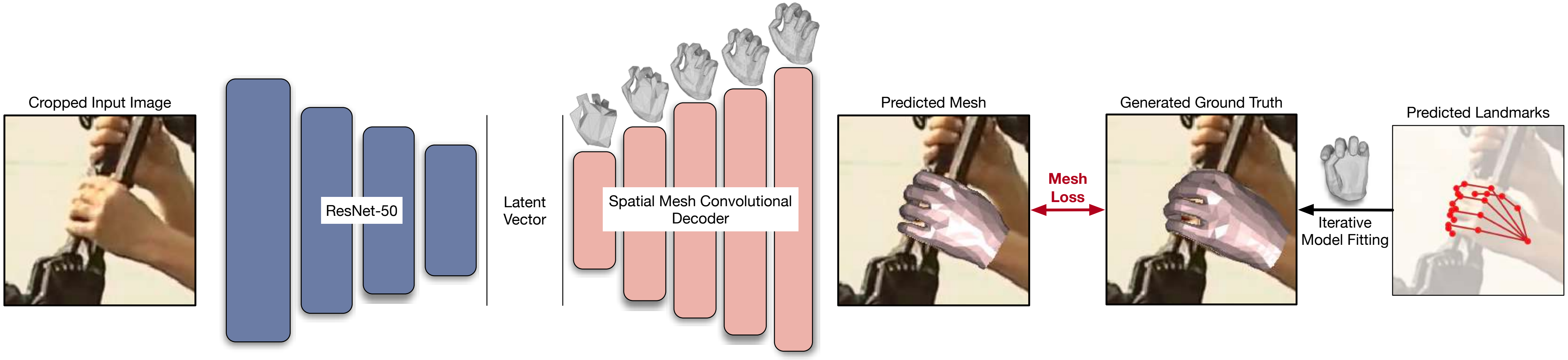
Learning 3D object models from 2D images



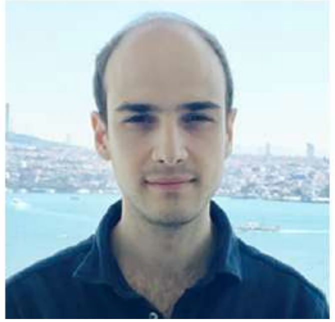
Learning from Imperfect Data Workshop

Iasonas Kokkinos

Ariel AI UCL

Ariel AI


G. Papandreou


R. A. Guler


B. Fulkerson


S. Zafeiriou


E. Schmitt


H. Wang


D. Kulon


P. Koutras


E. Skordos


S. Galanakis


A. Kakolyris


D. Stoddard


H. Tam


A. Lazarou

UCL, Imperial College, FAIR, INRIA, Stony Brook


M. Bronstein
Imperial College


Natalia Neverova
FAIR


Z. Shu
Stony Brook


M. Sahasrabudhe
INRIA


E. Bartrum
UCL


N. Paragios
INRIA


D. Samaras
Stony Brook

Human analysis: from coarse to fine

Image Classification

Input Image

Is there a person in this image?

Yes? **No?**

Image Classification

Human analysis: from coarse to fine

Person Detection

Input Image

Localize persons in the image.

Image Classification

Person Detection

Human analysis: from coarse to fine

Part Segmentation

Input Image

Segment semantically meaningful body parts.

Image Classification

Person Detection

Part Segmentation

Human analysis: from coarse to fine

Input Image

Pose Estimation

Localize joints of the persons in the images.

Image Classification

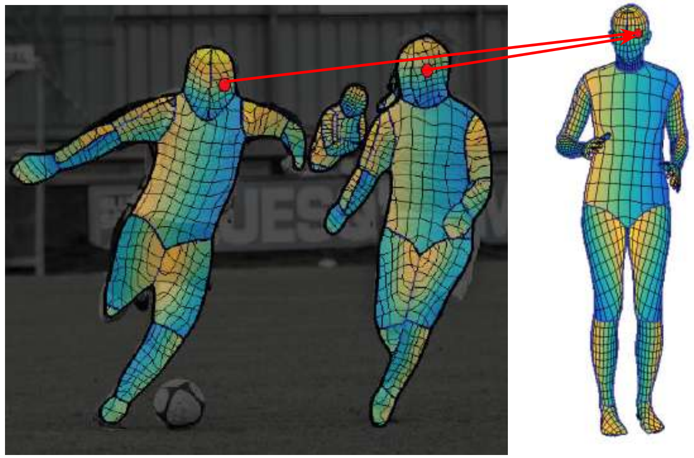
Person Detection

Part Segmentation

Pose Estimation

Human analysis: from coarse to fine

Dense Pose Estimation



Input Image

Find correspondence between all pixels and a 3D model.

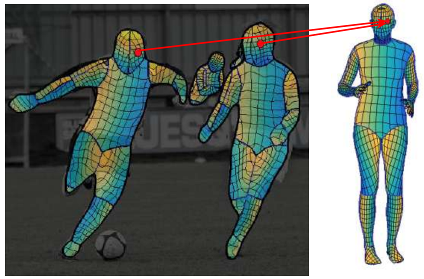
Image Classification

Person Detection

Part Segmentation

Pose Estimation

DensePose



Holy grail: 3D human reconstruction

“Wide Open”
(The Mill, 2015)

Ariel AI: 3D human reconstruction on mobile

Ariel AI: 3D human reconstruction on mobile

Seamless augmented reality

Universal motion capture

Holographic telepresence

Immersive gaming

Personalised, experiential retail

Kinetic learning

Challenges

Depth/height ambiguity

3D from 2D: fundamentally ill-posed problem

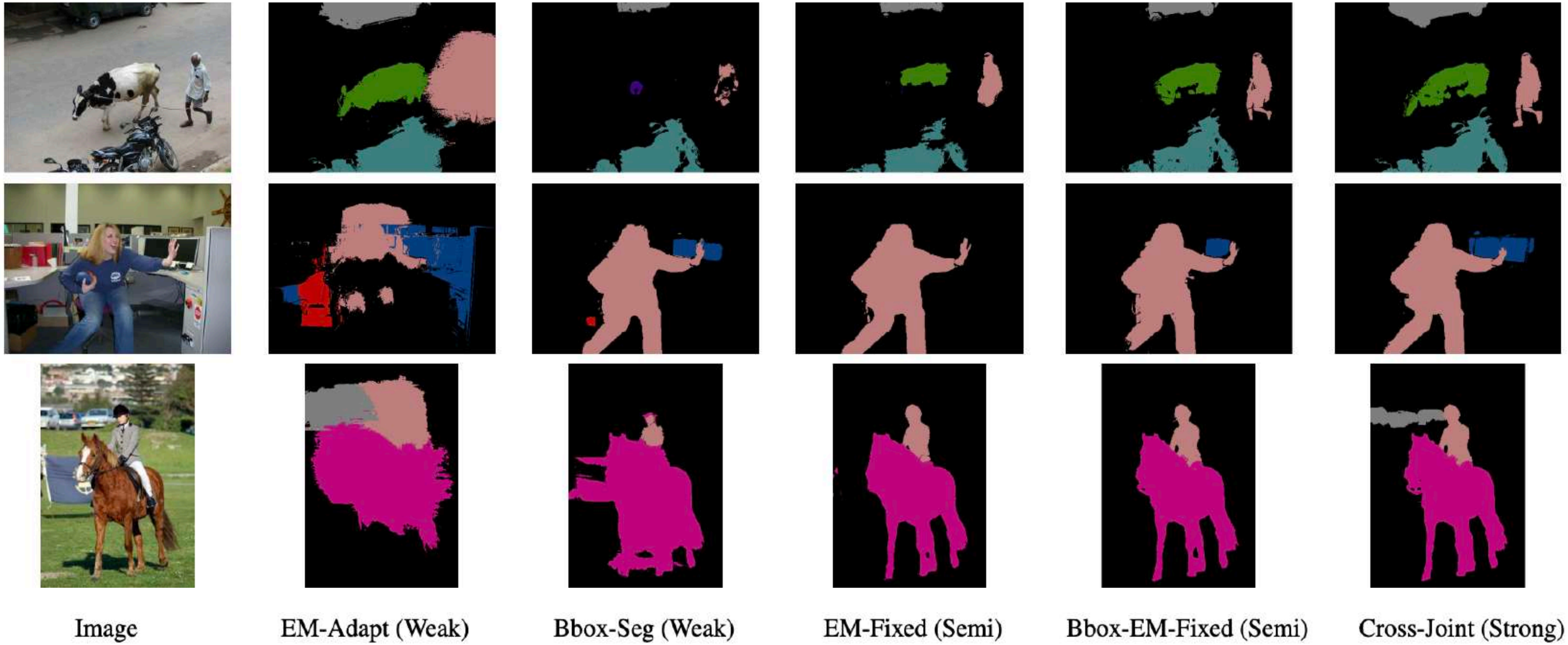
Scarce 3D supervision – almost impossible in-the-wild

From imperfect vision to imperfect data

Computer Vision before deep learning:
- Your `local evidence' is imperfect (classifier scores, unary terms, ..)
- Compensate for it by model-based prior during inference (AAMs, MRFs,..)

Computer Vision after deep learning:
- Your `local evidence' can become perfect
- Your training data is imperfect
- Compensate for it by some model-based prior, prior or during training

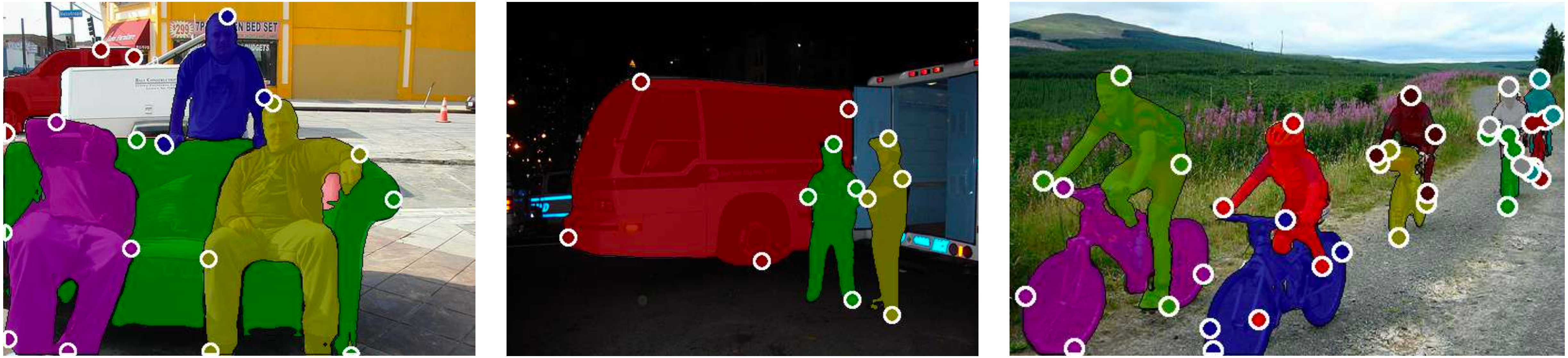
Imperfect Data for Semantic Segmentation



Bounding boxes + occupancy priors

“Weakly- and Semi-Supervised Learning of a Deep Convolutional Network for Semantic Image Segmentation” George Papandreou, Liang-Chieh Chen, Kevin P. Murphy, Alan L. Yuille, ICCV 2015

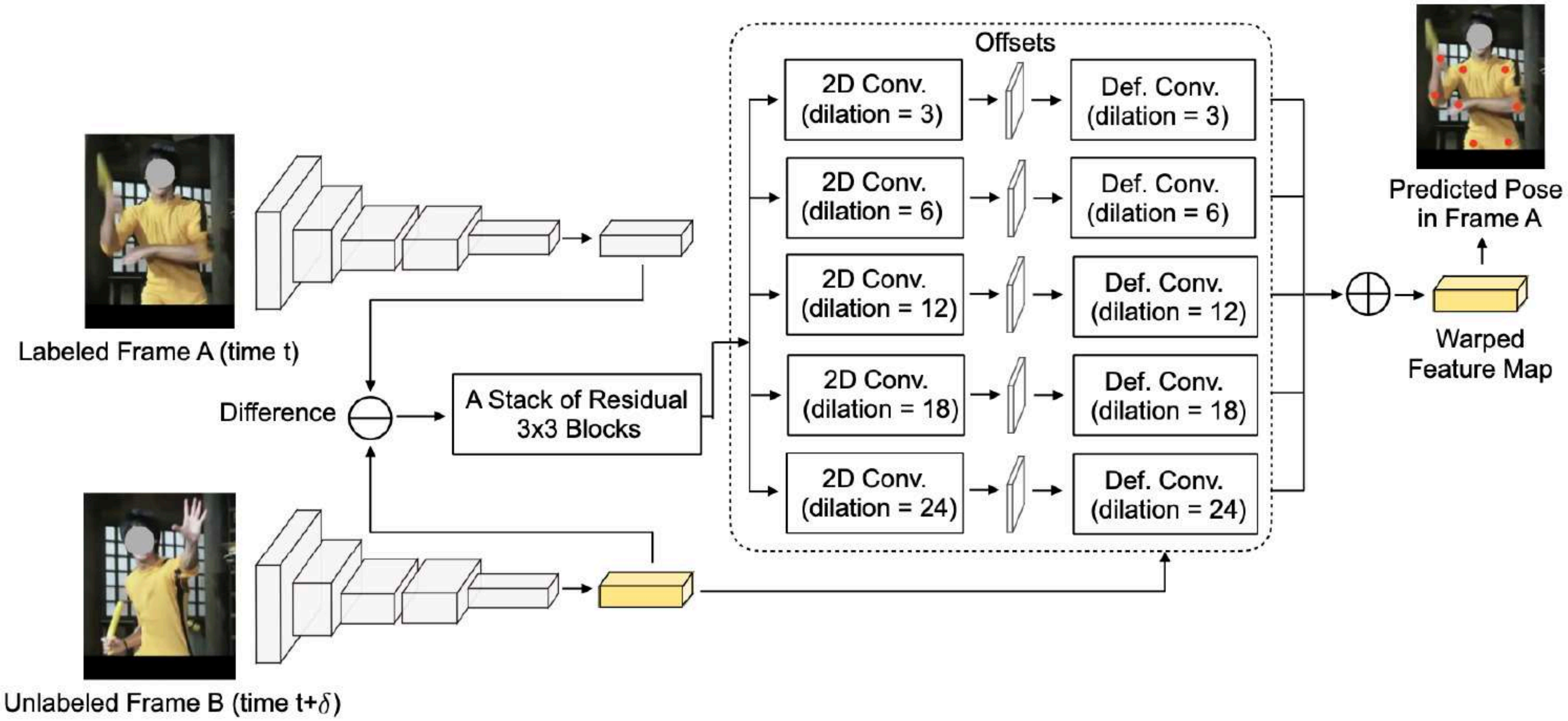
Imperfect Data for Instance Segmentation



4 points + segmentation system

Deep Extreme Cut: From Extreme Points to Object Segmentation,
Kevis-Kokitsi Maninis, Sergi Caelles, Jordi Pont-Tuset, Luc Van Gool

Imperfect Data for Pose Estimation



Keypoints + temporal correspondence

Learning Temporal Pose Estimation from Sparsely Labeled Videos, Bertasius, Gedas and Feichtenhofer, Christoph, and Tran, Du and Shi, Jianbo, and Torresani, Lorenzo(NeurIPS 2019)

Part 1: Weakly- and semi- supervised learning for 3D

HoloPose: Holistic 3D Human Reconstruction In-the-Wild, A. Guler and I. Kokkinos, CVPR 2019

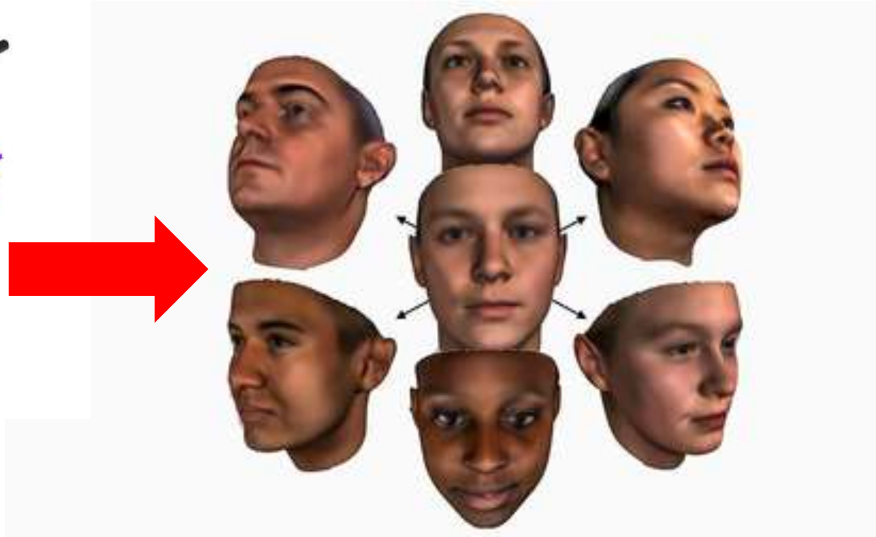
Weakly-Supervised Mesh-Convolutional Hand Reconstruction in the Wild, D. Kulon et al CVPR 2020

Part 2: Fully unsupervised learning for 3D

Unstructured face dataset

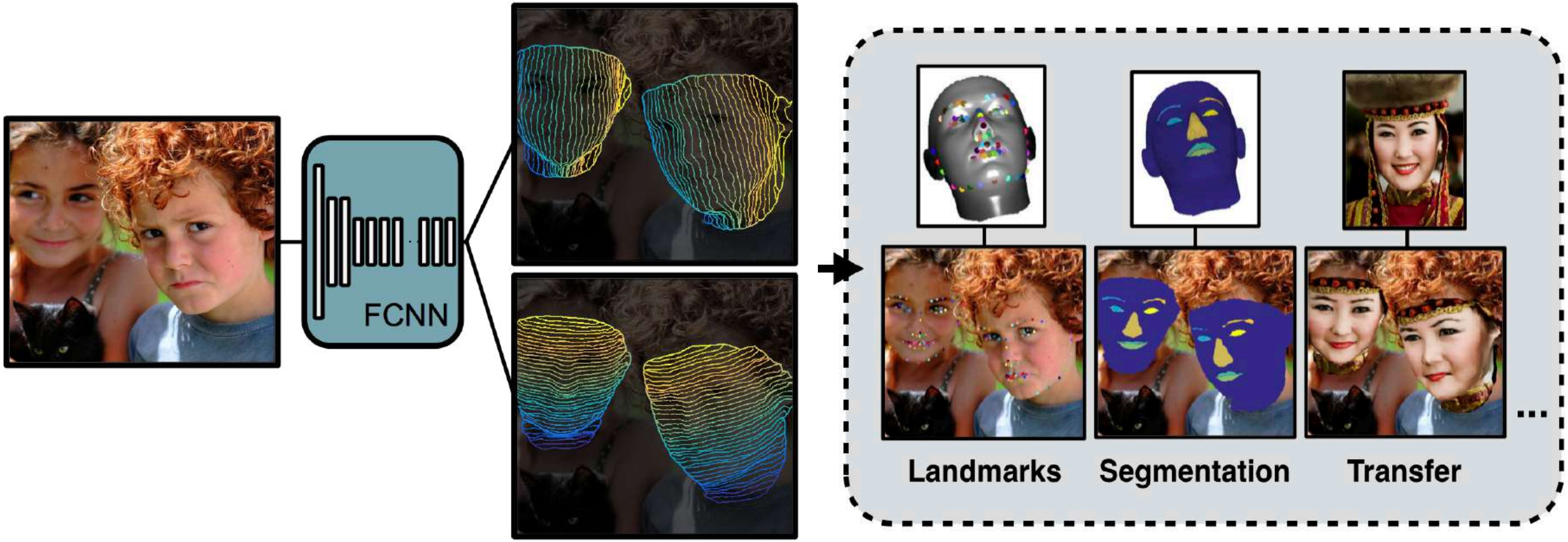
deep magic happens

3D model comes out



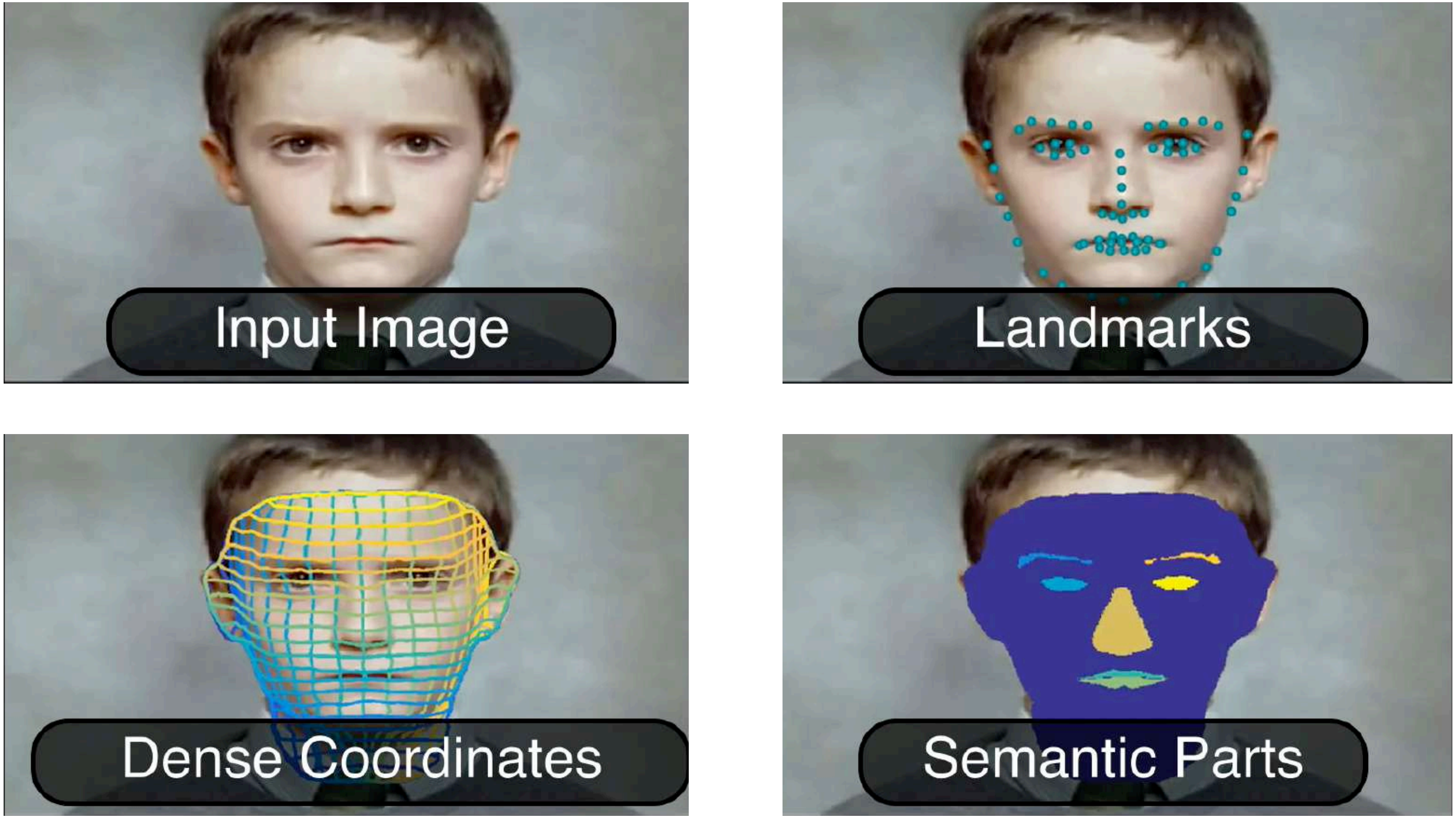
Lifting AutoEncoders: Unsupervised Learning of 3D Morphable Models Using Deep Non-Rigid Structure from Motion, M. Sahasrabudhe, Z. Shu, E. Bartrum, A. Guler, D. Samaras and I. Kokkinos, ICCV GMDL 2019

DenseReg: From Image to Template to Task



R. A. Guler, G. Trigeorgis, E. Antonakos, P. Snape, S. Zafeiriou, I. Kokkinos,
DenseReg: Fully Convolutional Dense Shape Regression In-the-Wild, CVPR 2017

DenseReg, Frame-by-Frame

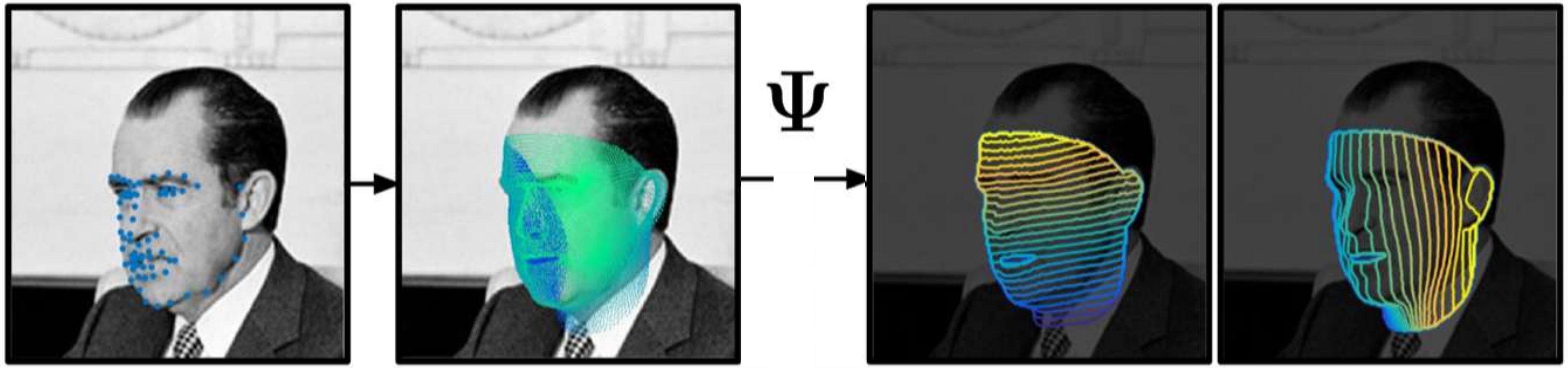


Supervision: from parametric model fitting to 2D keypoints

$$\boldsymbol{X} = [\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \ldots \boldsymbol{x}_m^\top]^\top \in \mathbb{R},$$

where each $\boldsymbol{x}_j \in \mathbb{R}^3$ is a vertex

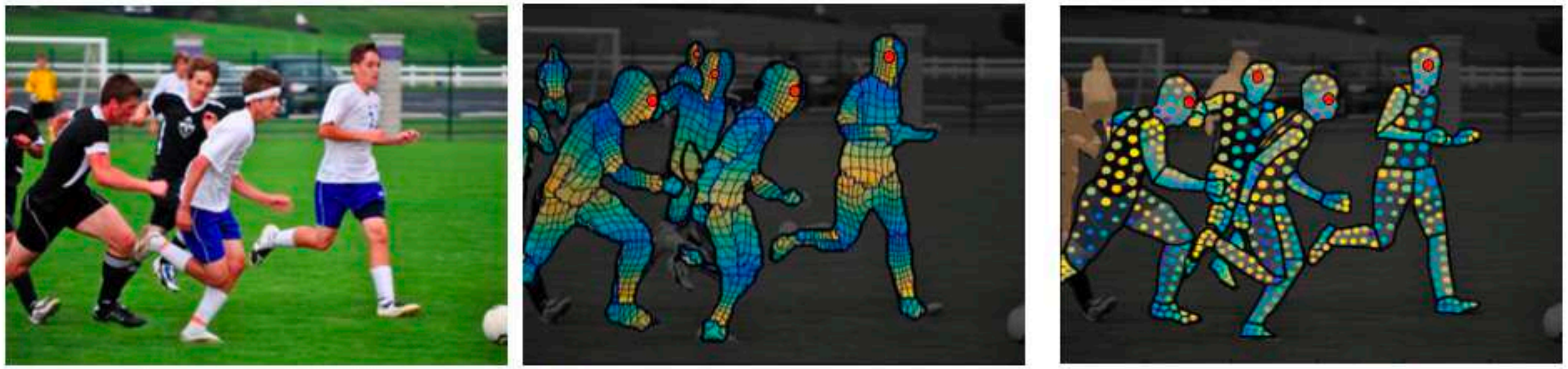
2D canonical coordinates $\boldsymbol{U} \in \mathbb{R}^{2\times m}$



Annotation effort: a few 2D landmarks per image

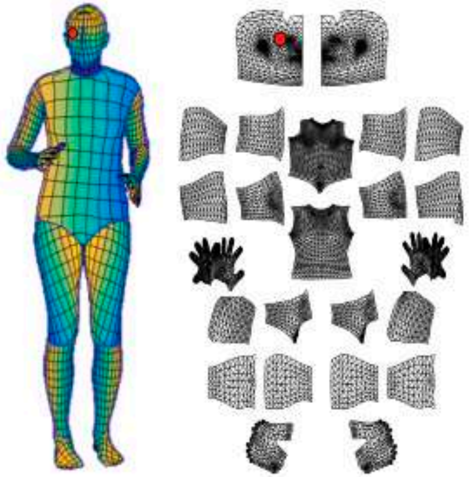
Density: morphable model prior

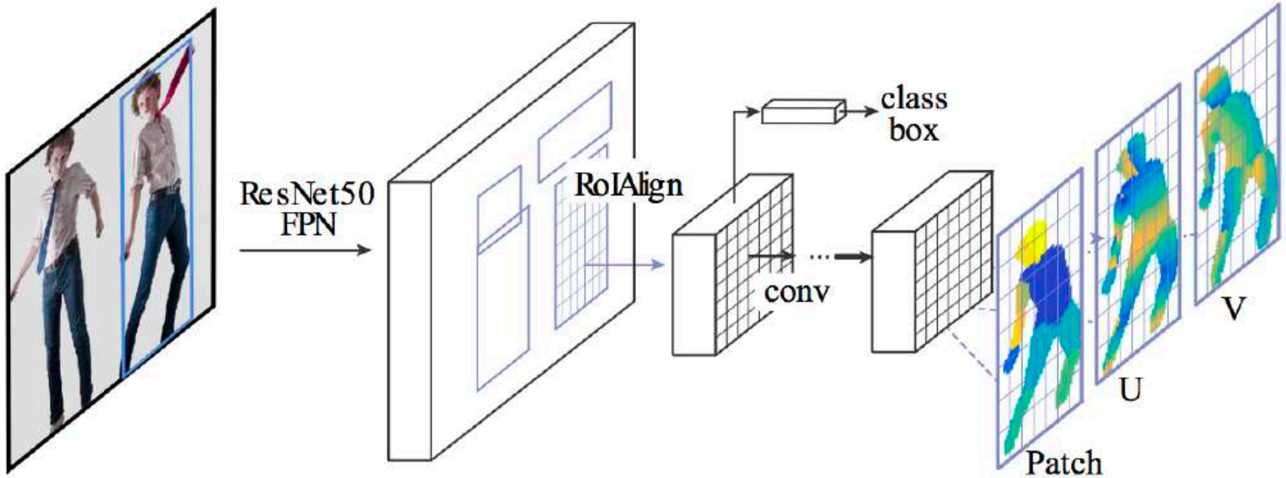
DensePose: dense image-to-body correspondence



DensePose-RCNN Results

DensePose COCO Dataset



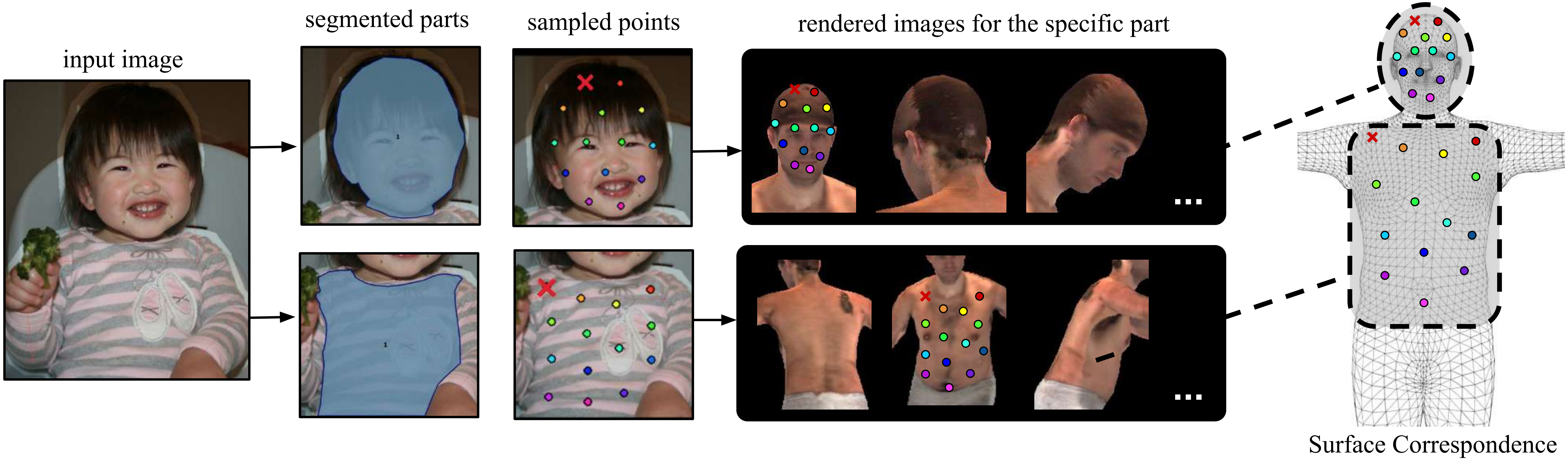


DensePose-RCNN: ~25 FPS

http://densepose.org/

R. A. Guler, N. Neverova, I. Kokkinos “DensePose: Dense Human Pose Estimation In The Wild”, CVPR’18

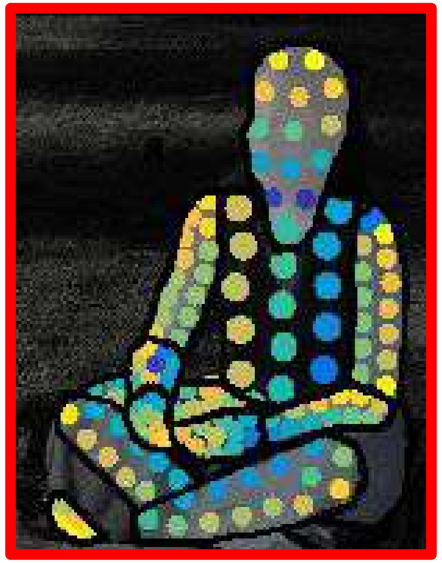
Annotation pipeline-II



DensePose-COCO dataset

densepose.org

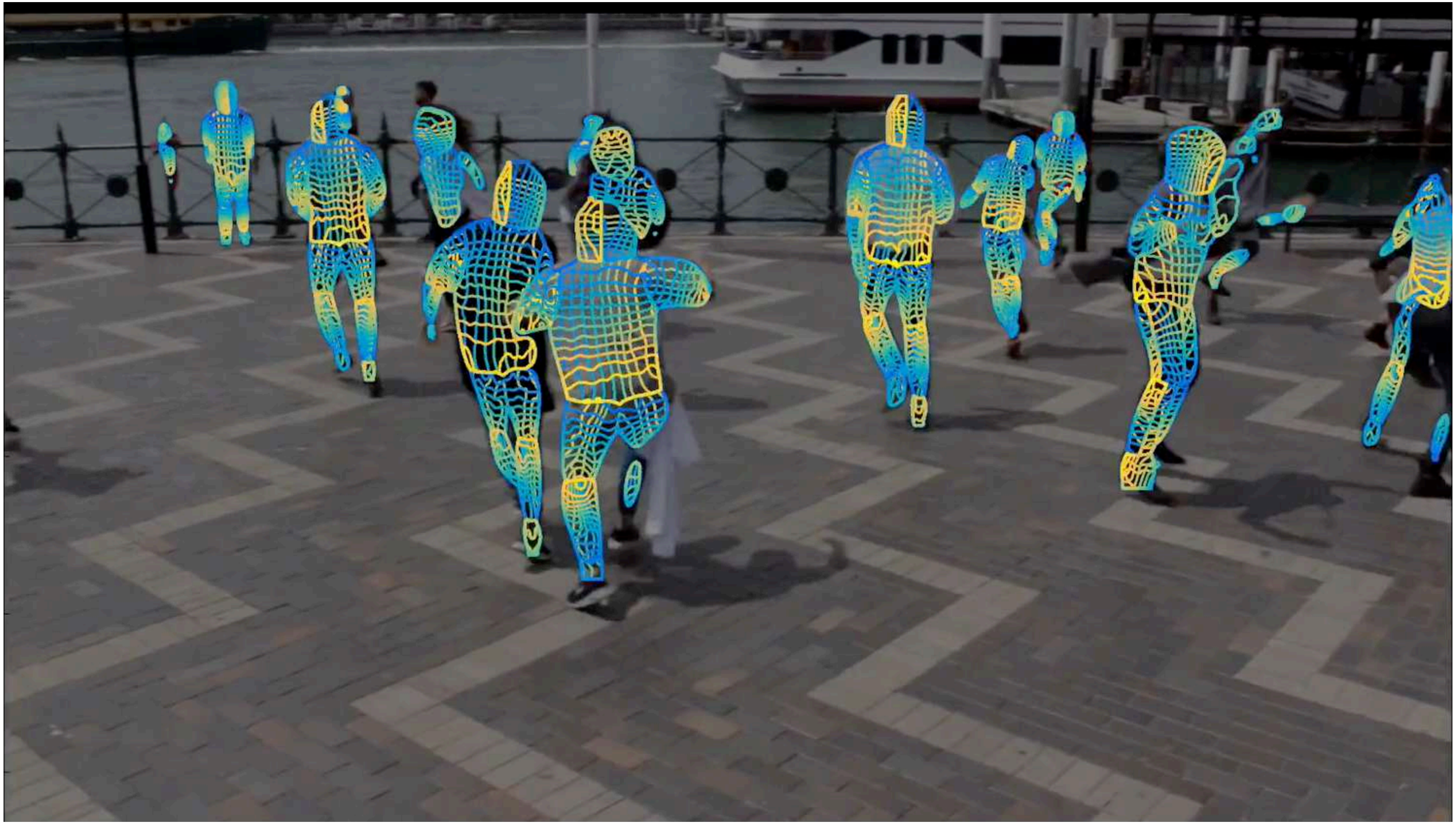
Image



U coordinates

V coordinates

DensePose-RCNN in action

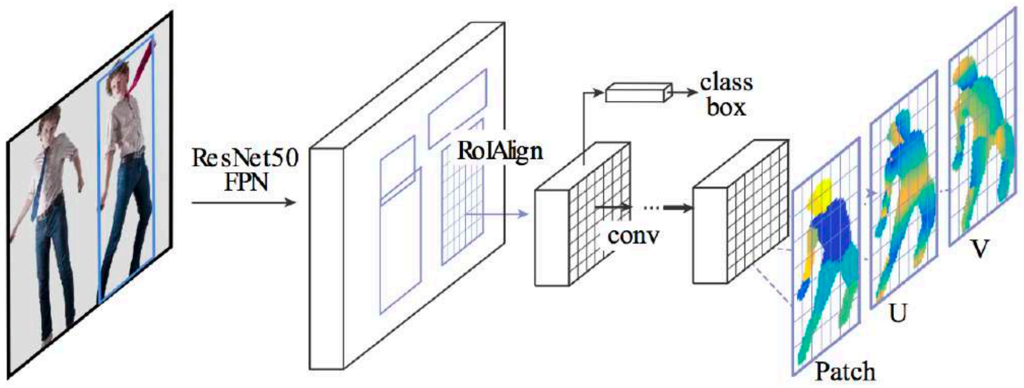


HoloPose: multi-person 3D reconstruction results

R. A. Guler, I. Kokkinos “HoloPose: Holistic 3D Human Reconstruction In The Wild”, CVPR’19

Surface-level human understanding, CVPR 2018

Dense UV coordinate regression

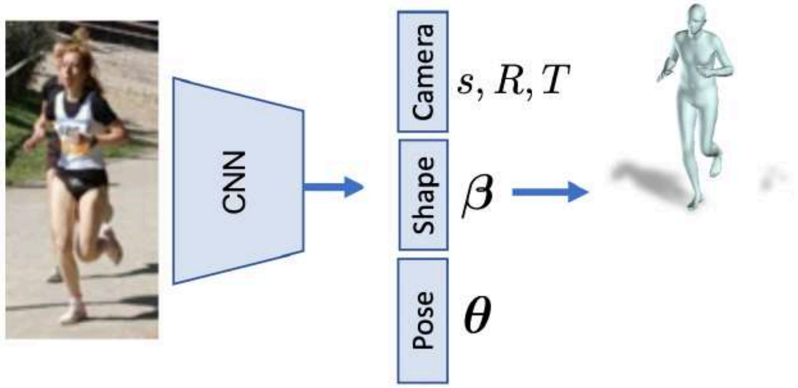


DensePose: Dense Human Pose Estimation In The Wild, CVPR 2018
R. A. Güler, N. Neverova, I. Kokkinos,

Robust & accurate, "in-the-wild"

Not 3D

SMPL parameter regression

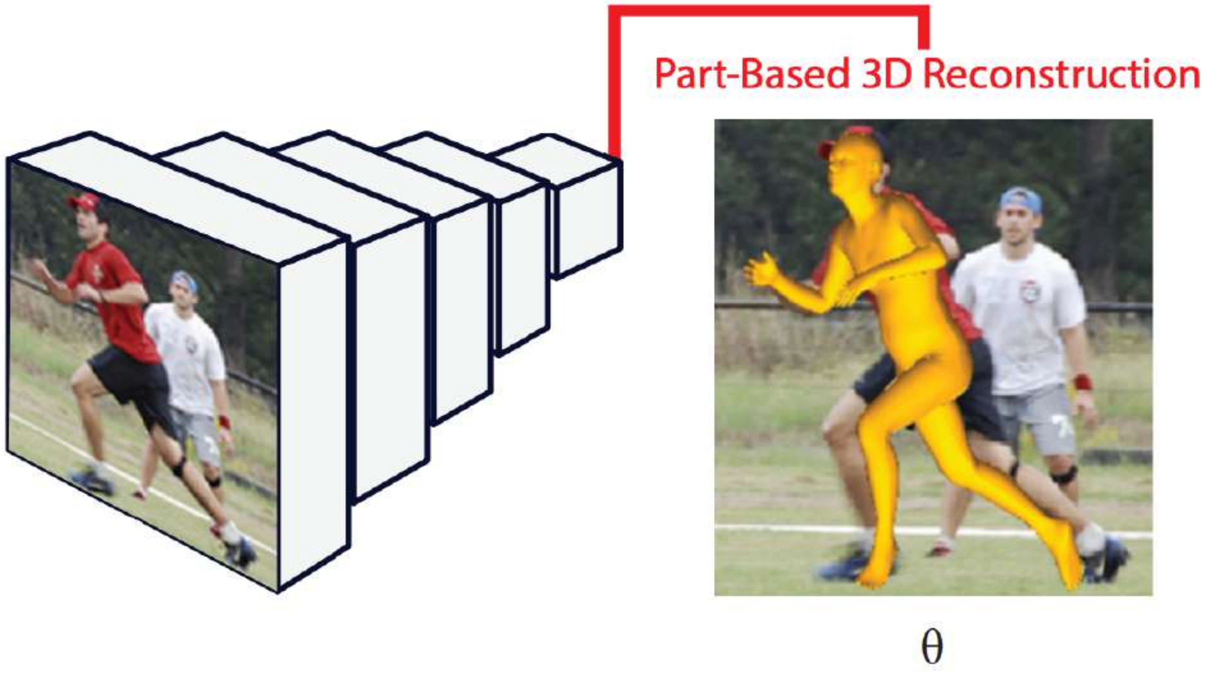


End-to-end Recovery of Human Shape and Pose, CVPR 2018
A. Kanazawa M. J Black D. W. Jacobs J. Malik
Learning to Estimate 3D Human Pose and Shape from a Single Image, CVPR 2018
G. Pavlakos, L. Zhu, X. Zhou, K. Daniilidis
Monocular 3D Pose and Shape Estimation of Multiple People, CVPR 2018,
Andrei Zanfir, Elisabeta Marinoiu, Cristian Sminchisescu

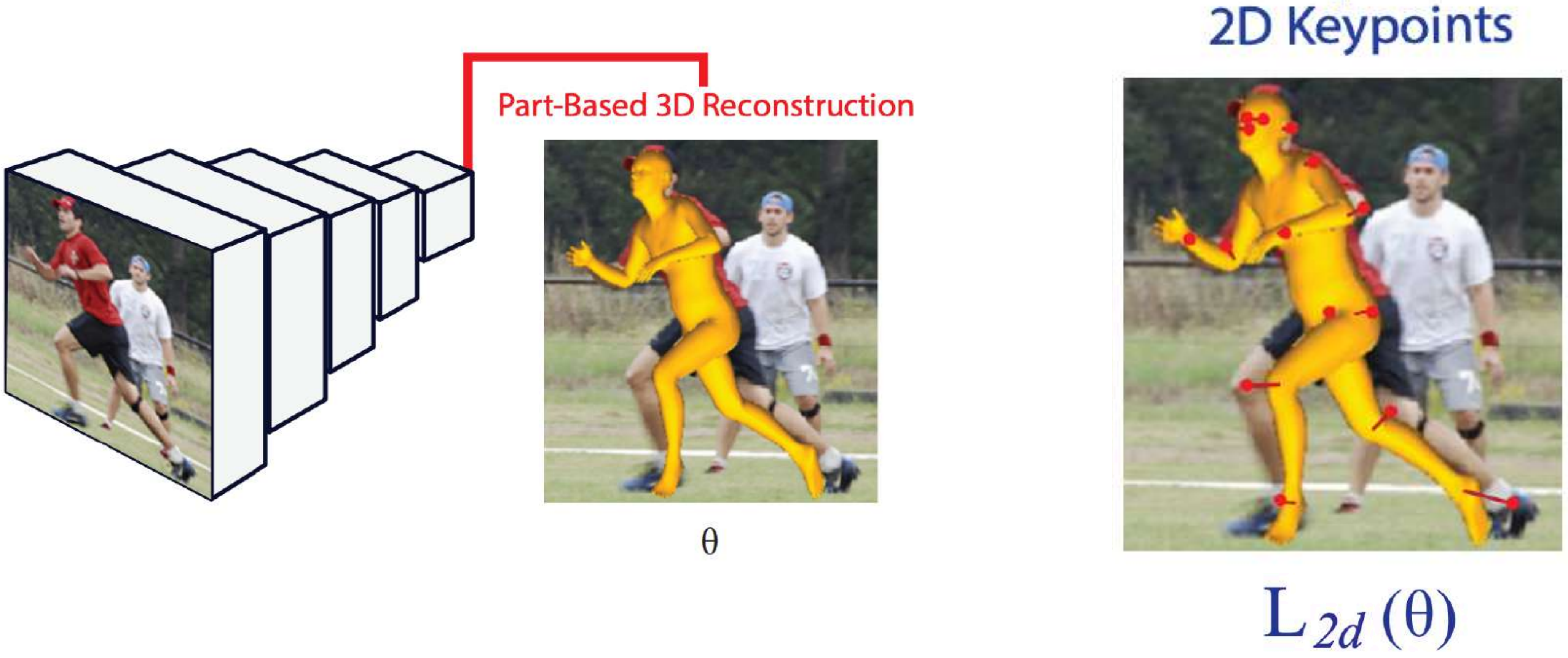
Parametric and 3D

Alignment

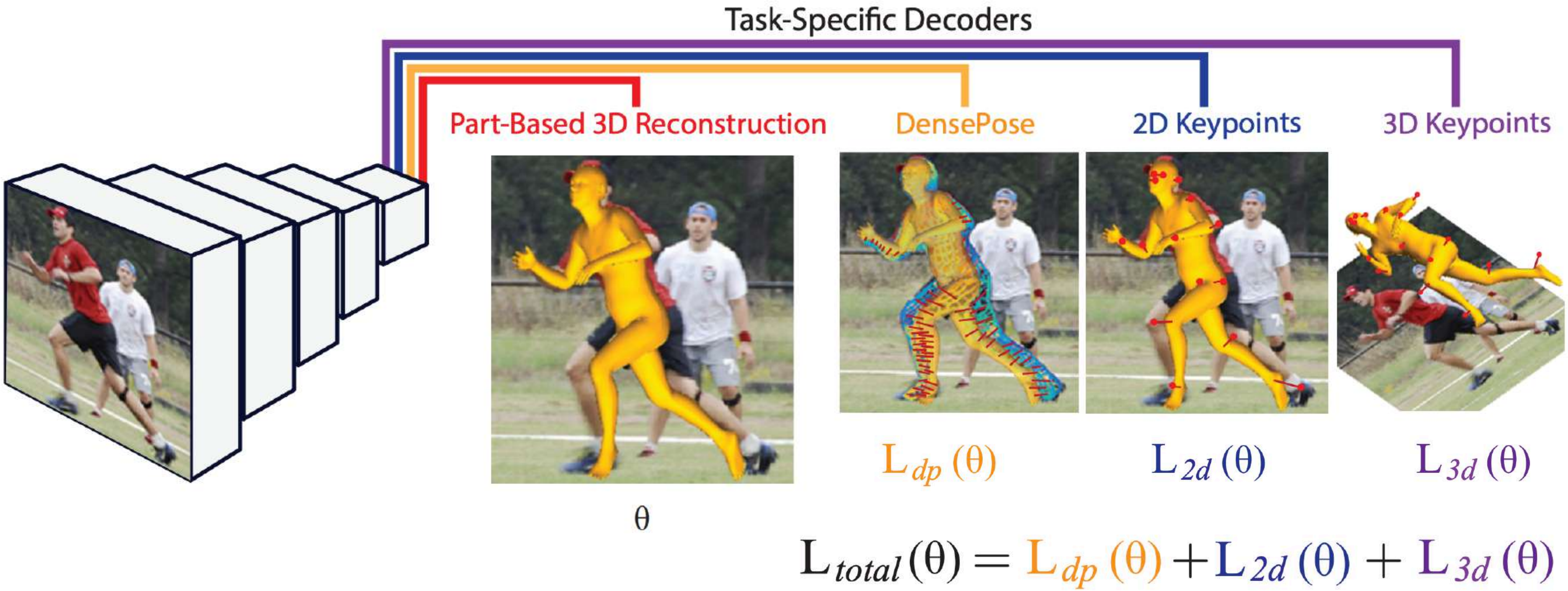
Bottom-up human body reconstruction



Bottom-up 2D Keypoint localization

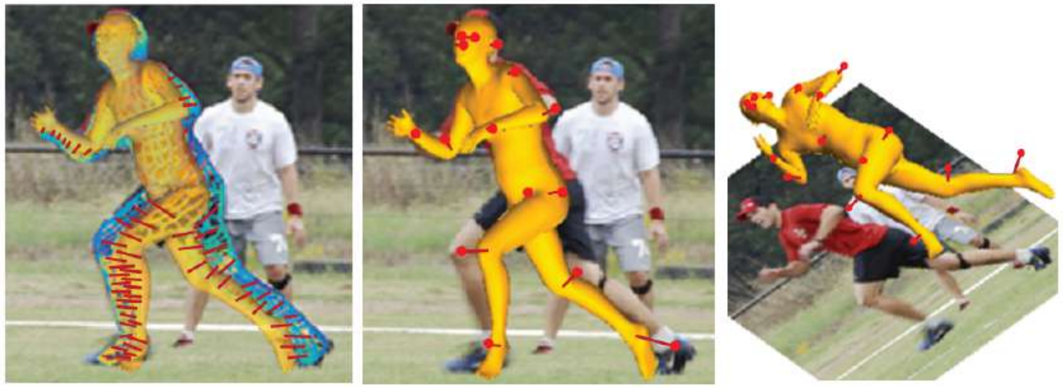


Bottom-up/Top-down Synergistic Refinement



Synergistic Refinement

DensePose | 2D Keypoints | 3D Keypoints



$L_{dp}(\theta)$ $L_{2d}(\theta)$ $L_{3d}(\theta)$

$$L_{total}(\theta) = L_{dp}(\theta) + L_{2d}(\theta) + L_{3d}(\theta)$$

Before/After Refinement | Final Result

$\theta, \ \theta^*$

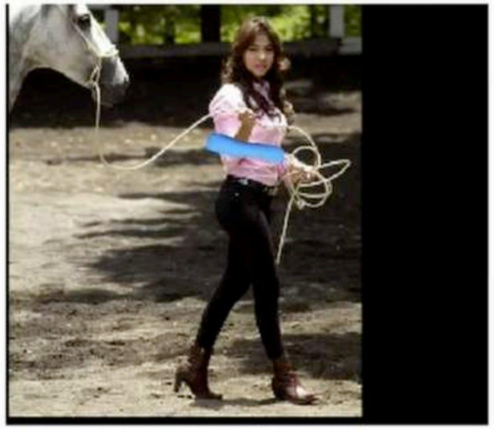
$$\theta^* = \underset{\theta}{\operatorname{argmin}} \ L_{total}(\theta)$$

3D Pose Estimation Results

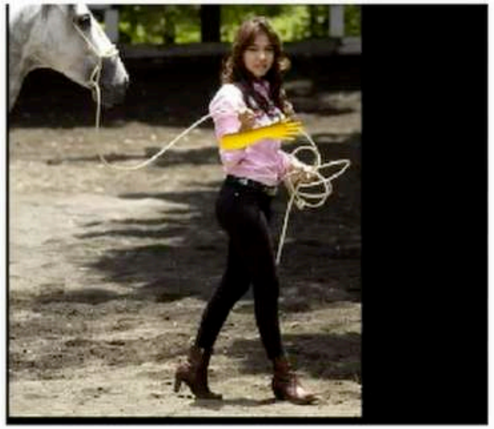
Human 3.6m Dataset

Method	*PA MPJPE*	*MPJPE*
HMR	56.8	87.97
Ours	**50.56**	**64.28**
Ours+ Synergy	**46.52**	**60.27**

HMR
Kanazawa et al. CVPR 2018



Ours
Part-Based 3D Reconstruction

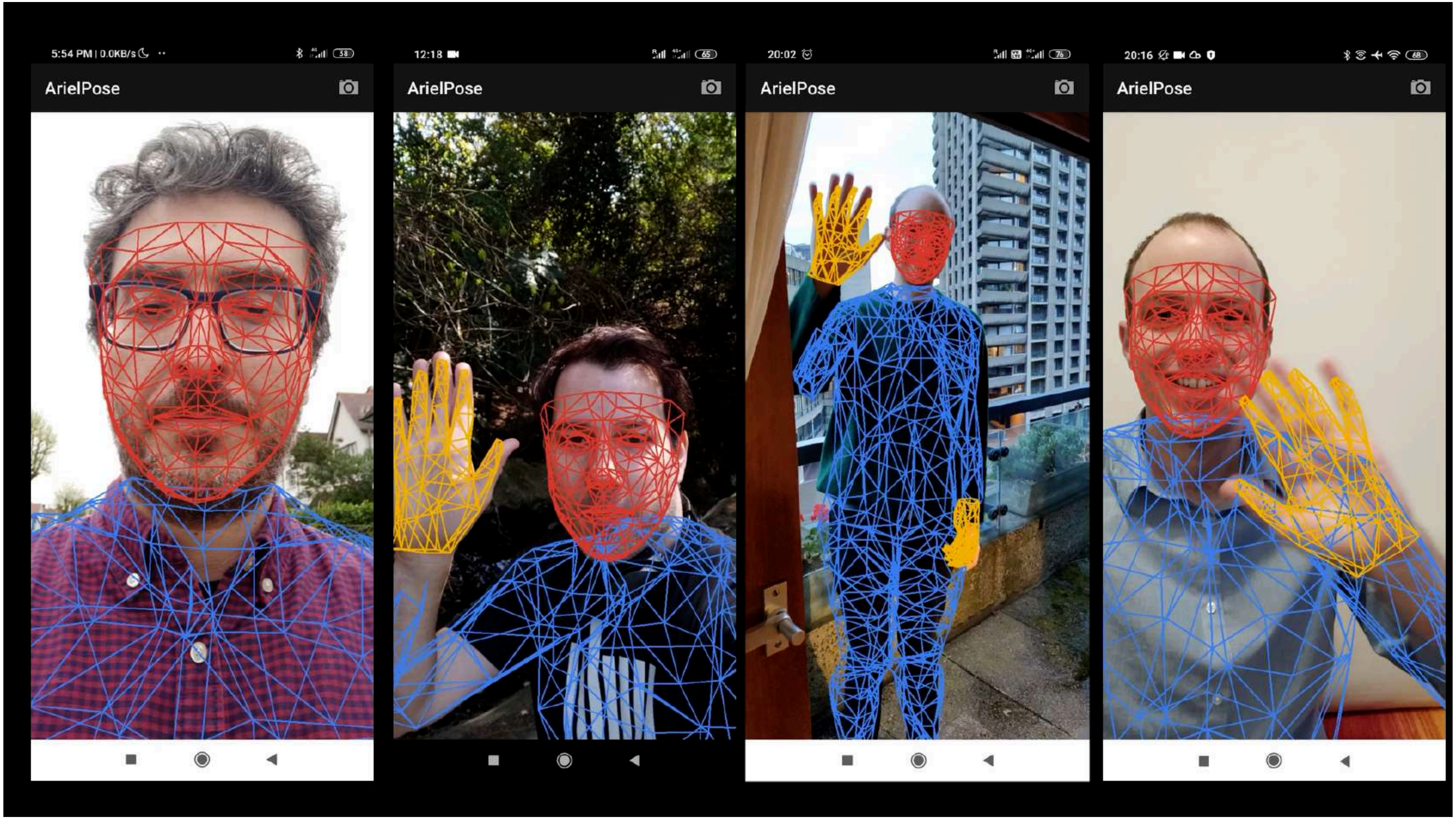


Ours
Synergistic Refinement

Ariel Holopose 2019

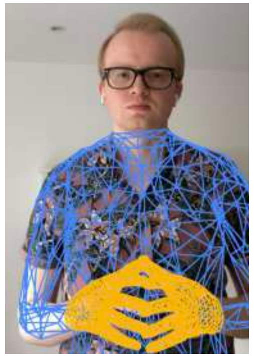
- In-the-wild human 3D reconstruction

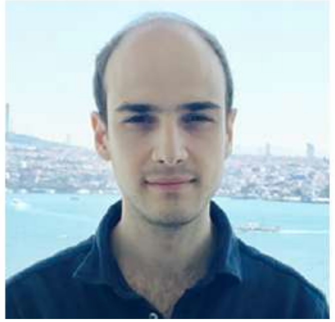
Ariel Holopose 2020



Weakly-Supervised Mesh-Convolutional Hand Reconstruction in the Wild

Dominik Kulon Riza Alp Güler Iasonas Kokkinos Michael Bronstein Stefanos Zafeiriou





arielai.com/mesh_hands

Oral, CVPR 2020

Poster, Fourth Workshop on Computer Vision for AR/VR

Motivation - hand pose estimation

youtu.be/aQ4shIsQabo

- **Broad array of applications:**
  - human-computer interaction
  - augmented reality
  - virtual telepresence
  - sign language recognition
- **Existing approaches do not always:**
  - Generalize to non-laboratory environments.
  - Provide full mesh reconstruction.
  - Operate in real time.

Hand Reconstruction System

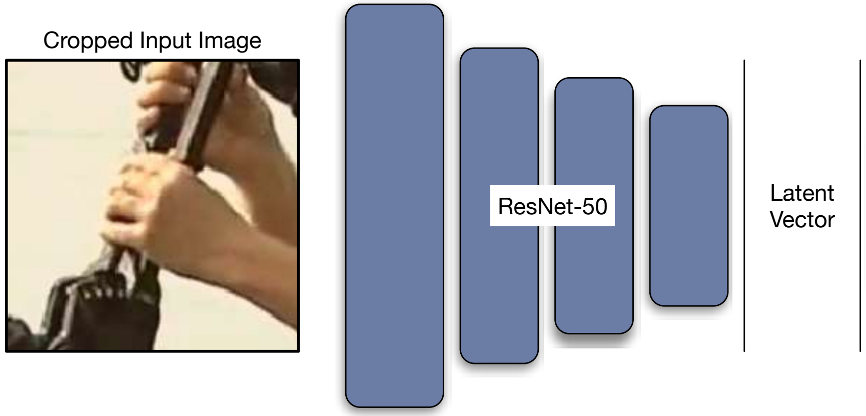
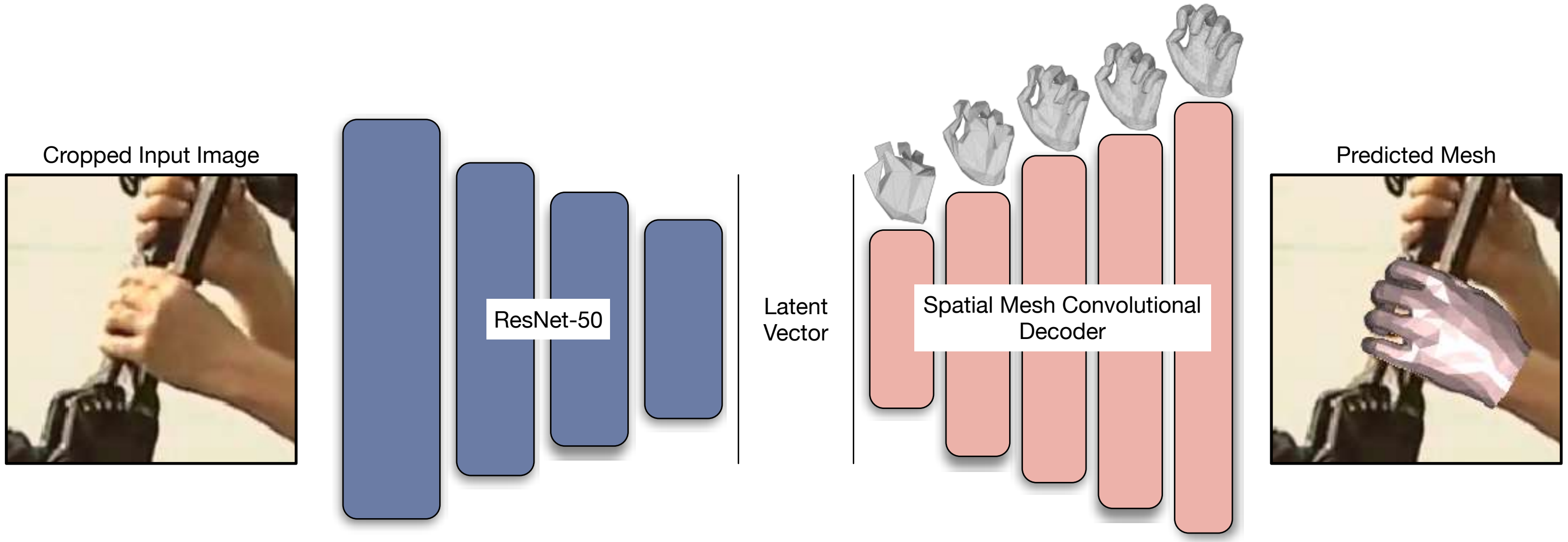


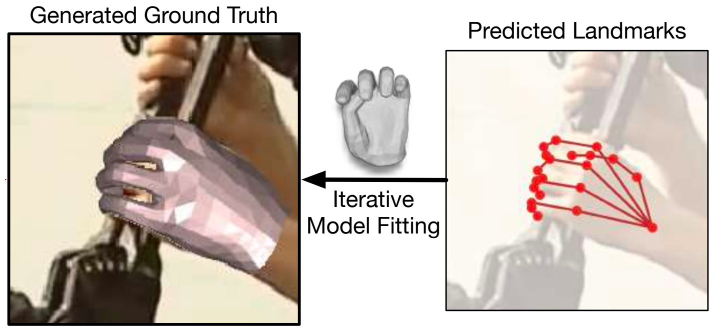
Image Encoding

Hand Reconstruction System



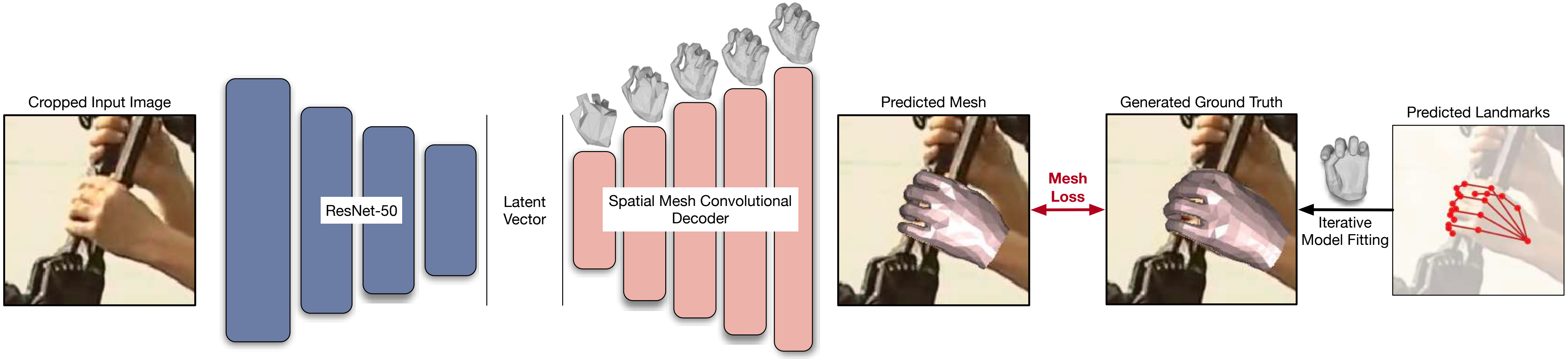
Mesh Reconstruction

Hand Reconstruction System



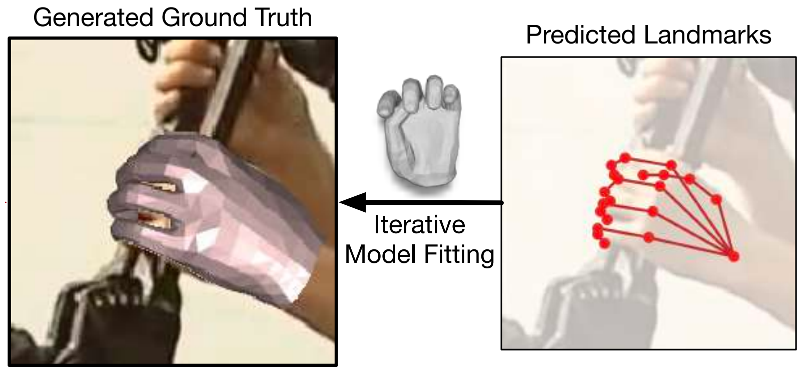
Weak Supervision

Hand Reconstruction System



End-to-End Training

Parametric Hand Model Fitting



$$\{\beta^*, w^*, \vec{T}_\delta^*, s^*\} = \arg \min_{\beta, w, \vec{T}_\delta, s} (E_{2D} + E_{bone} + E_{reg})$$

- **2D Reprojection Term**

  Minimizes the distance between 2D joints.

- **Bone Length Preservation Term**

  Ensures that the length of each edge in the hand skeleton tree is preserved.

- **Regularization Term**

  Penalizes deviations from the mean pose.

- **K-Means Prior**

  We constrain joint angles to lie in the convex hull of pre-computed cluster centers.

Novel Dataset

We release a dataset of meshes aligned with in the wild images.

- Training set: 102 videos.
- Validation and test sets: 7 videos.
- Hundreds of subjects.
- 50K samples.


u reach for me
acquired throu

Evaluation - standard benchmarks

We also obtain state-of-the-art performance on popular laboratory datasets.

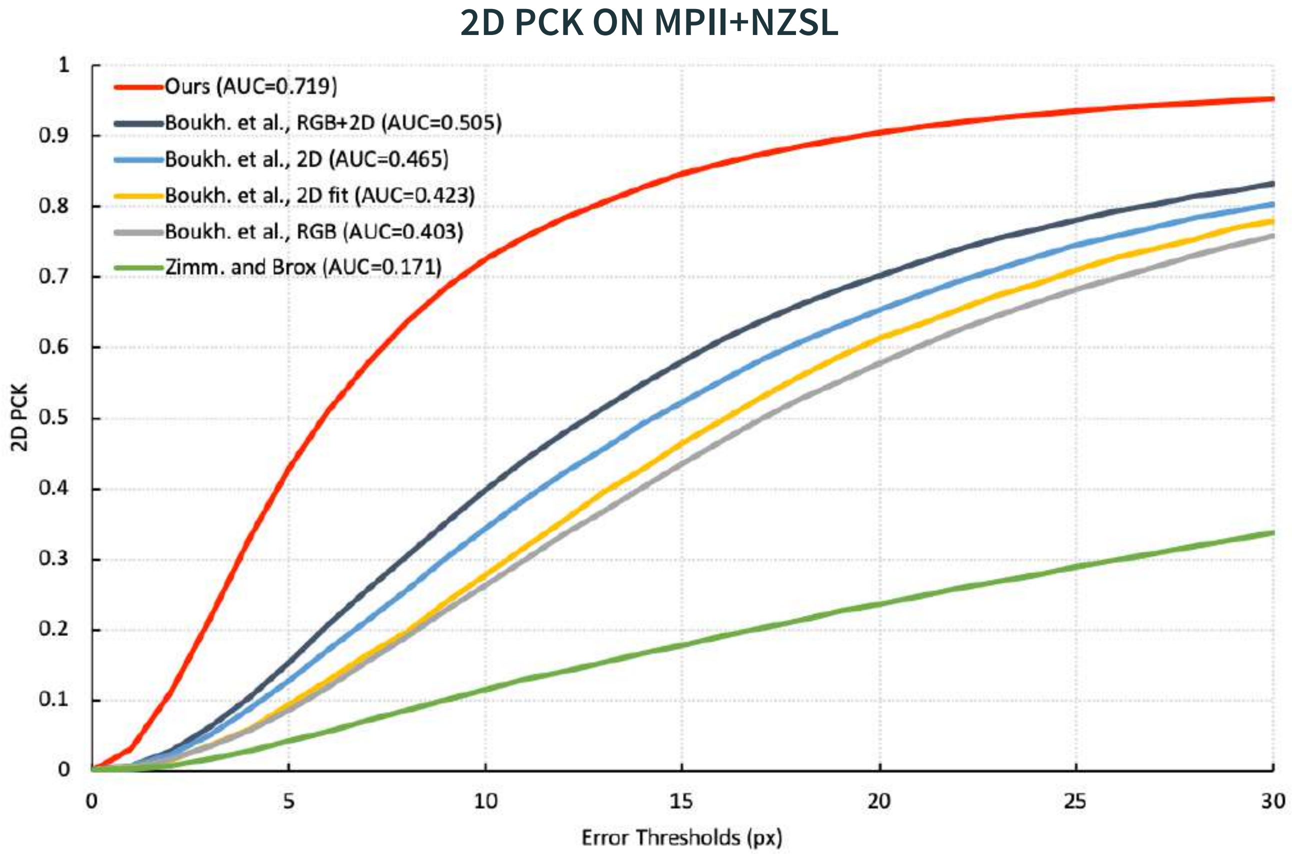
- Rendered Handpose Dataset (RHD)
- FreiHAND

Method (synthetic, 3D)	RHD (AUC)	Mesh	Speed (FPS)
Zimm. and Brox (2017)	0.675		
Yang and Yao (2019)	0.849		
Spurr et al. (2018)	0.849		
Zhou et al. (2020)	0.856		100 (GPU)
Cai et al. (2018)	0.887		
Zhang et al. (2019)	0.901		
Ge et al. (2019)	0.92		50 (GPU)
Baek et al. (2019)	0.926		
Yang et al. (2019)	0.943		
Ours	**0.956**		70 (GPU)

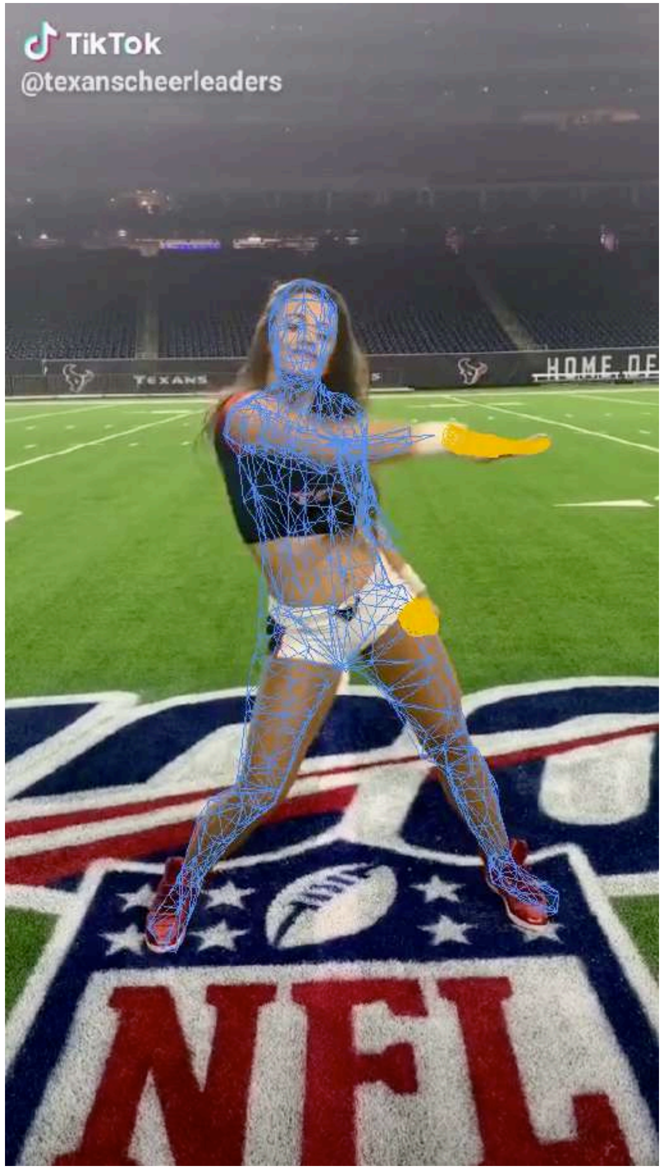
Evaluation - in the wild

We largely outperform other approaches on an in the wild benchmark.

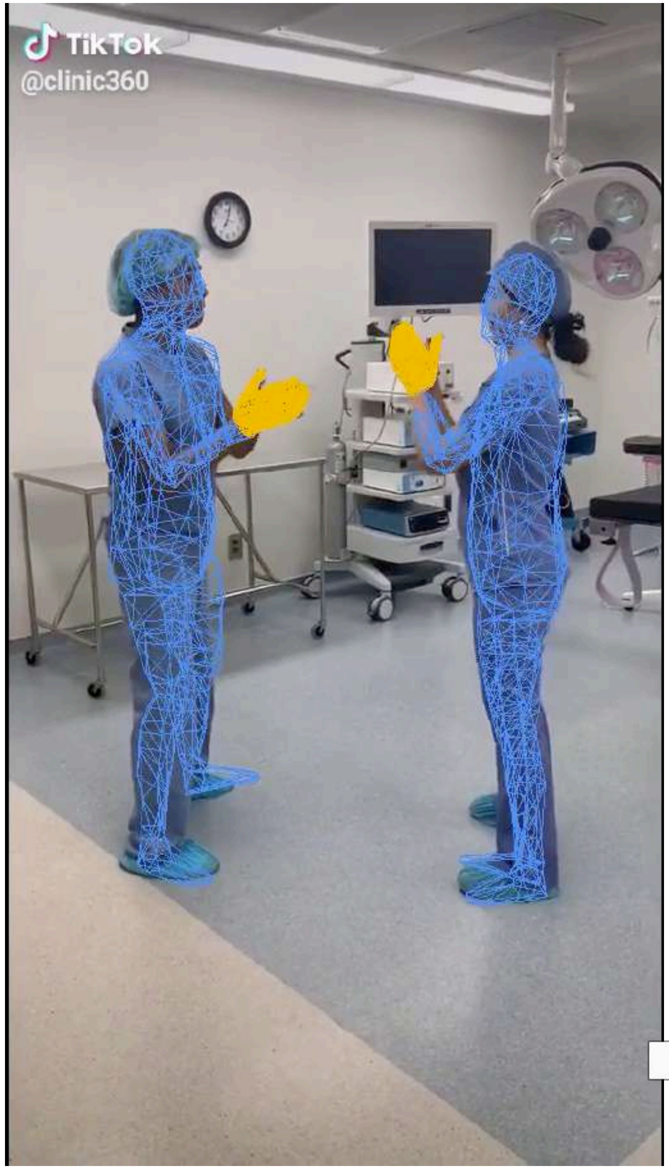
MPII+NZSL Dataset



Ariel AI


TikTok
@texanscheerleaders
TEXANS
HOME OF
NFL


TikTok
@mattsteffanina
FILA


TikTok
@clinic360

Egocentric Perspective

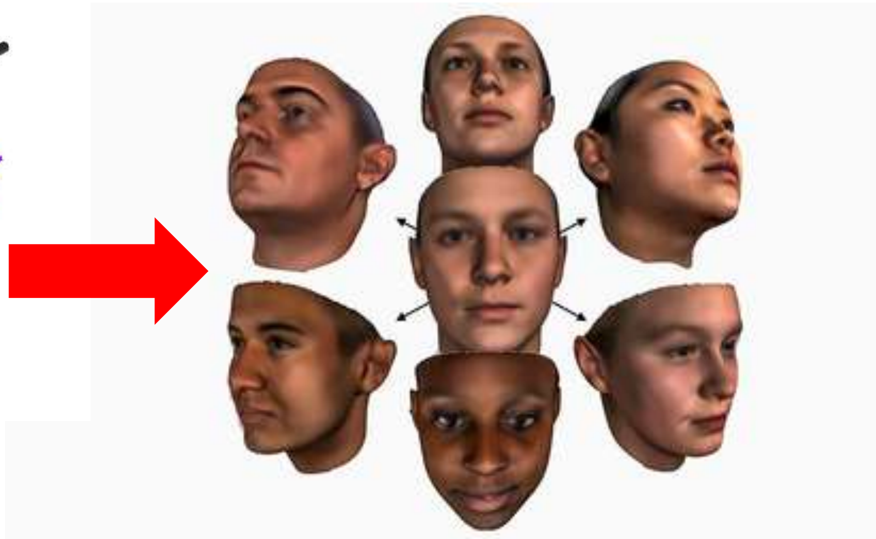
AR Effects

Part 2: Lifting AutoEncoders: Unsupervised 2D-to-3D

Unstructured face dataset

deep magic happens

3D model comes out



Lifting AutoEncoders: Unsupervised Learning of 3D Morphable Models Using Deep Non-Rigid Structure from Motion, M. Sahasrabudhe, Z. Shu, E. Bartrum, A. Guler, D. Samaras and I. Kokkinos, Arxiv 2019

Unsupervised learning of deformable models

Learning a template and the deformation for a class of images.

A canonical
appearance template

A class of images (MNIST 3)

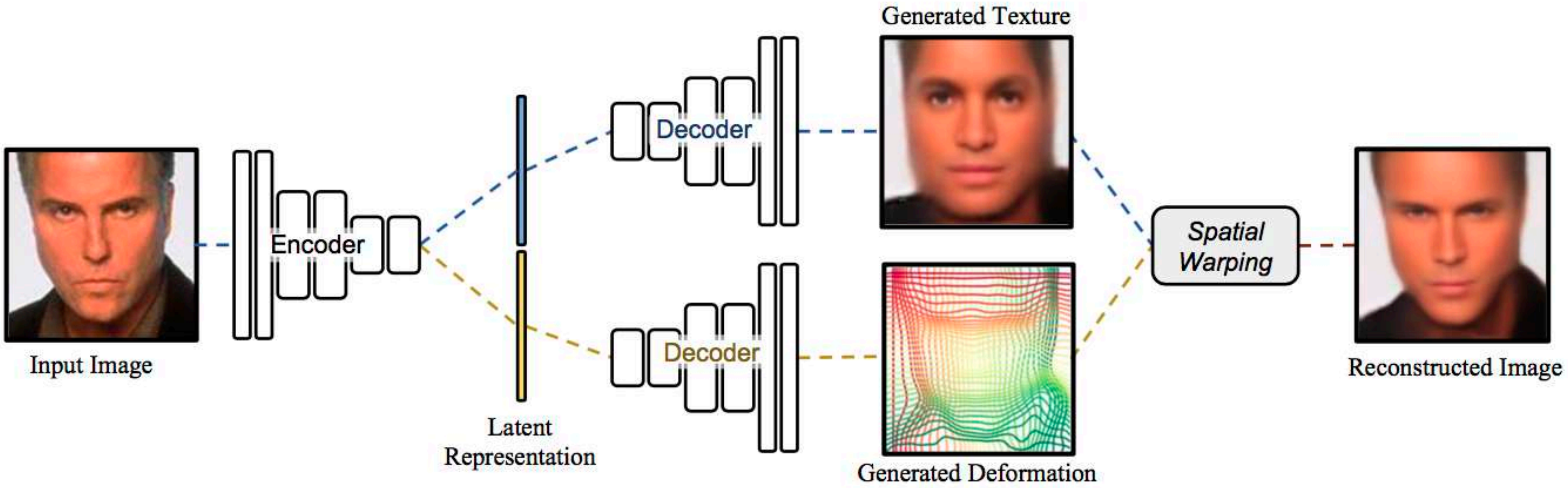
Unsupervised learning of deformable models

Goal: learn a template and the deformation for a class of images.

A canonical
appearance template

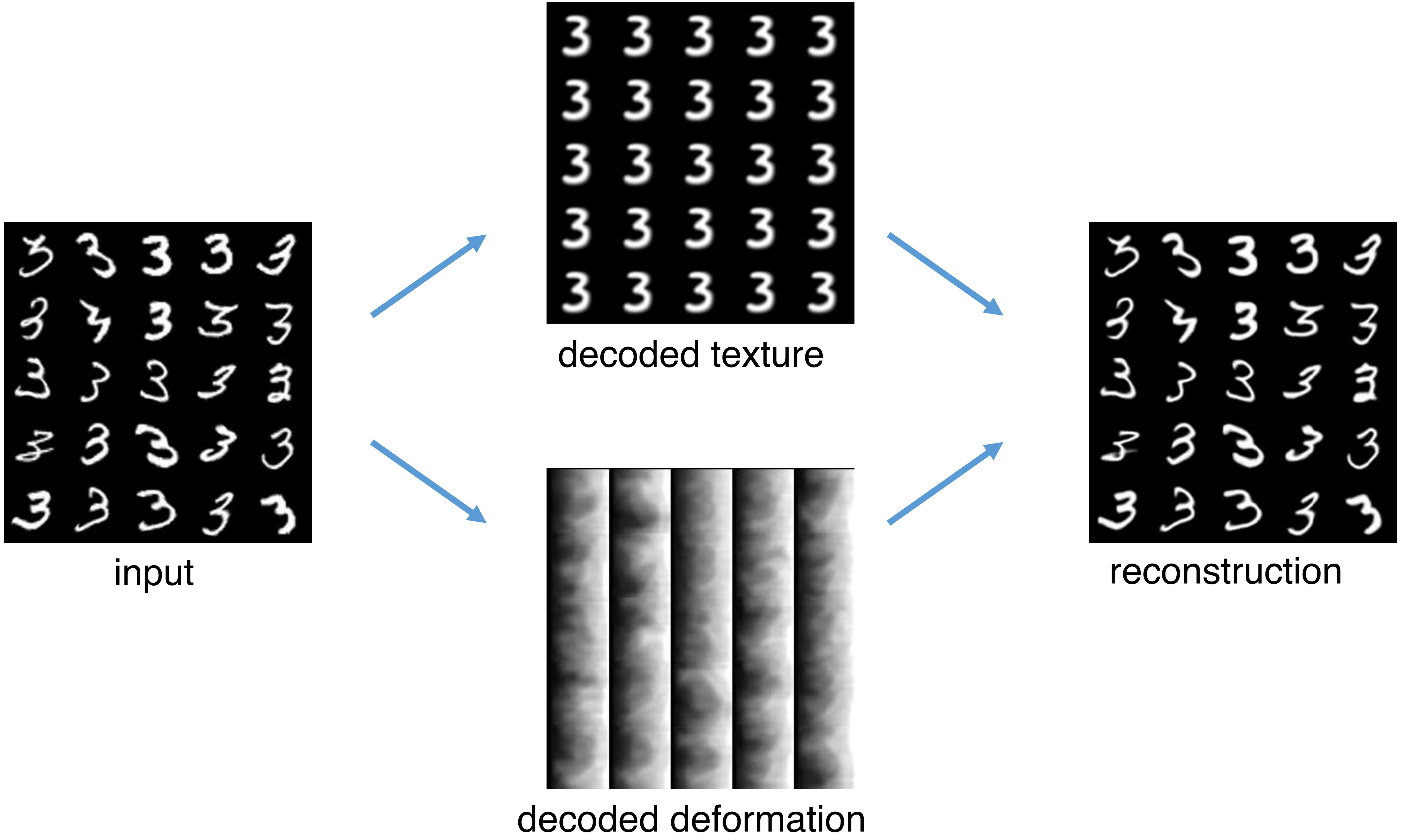
A class of images (Faces)

Deforming AutoEncoder (DAE) model



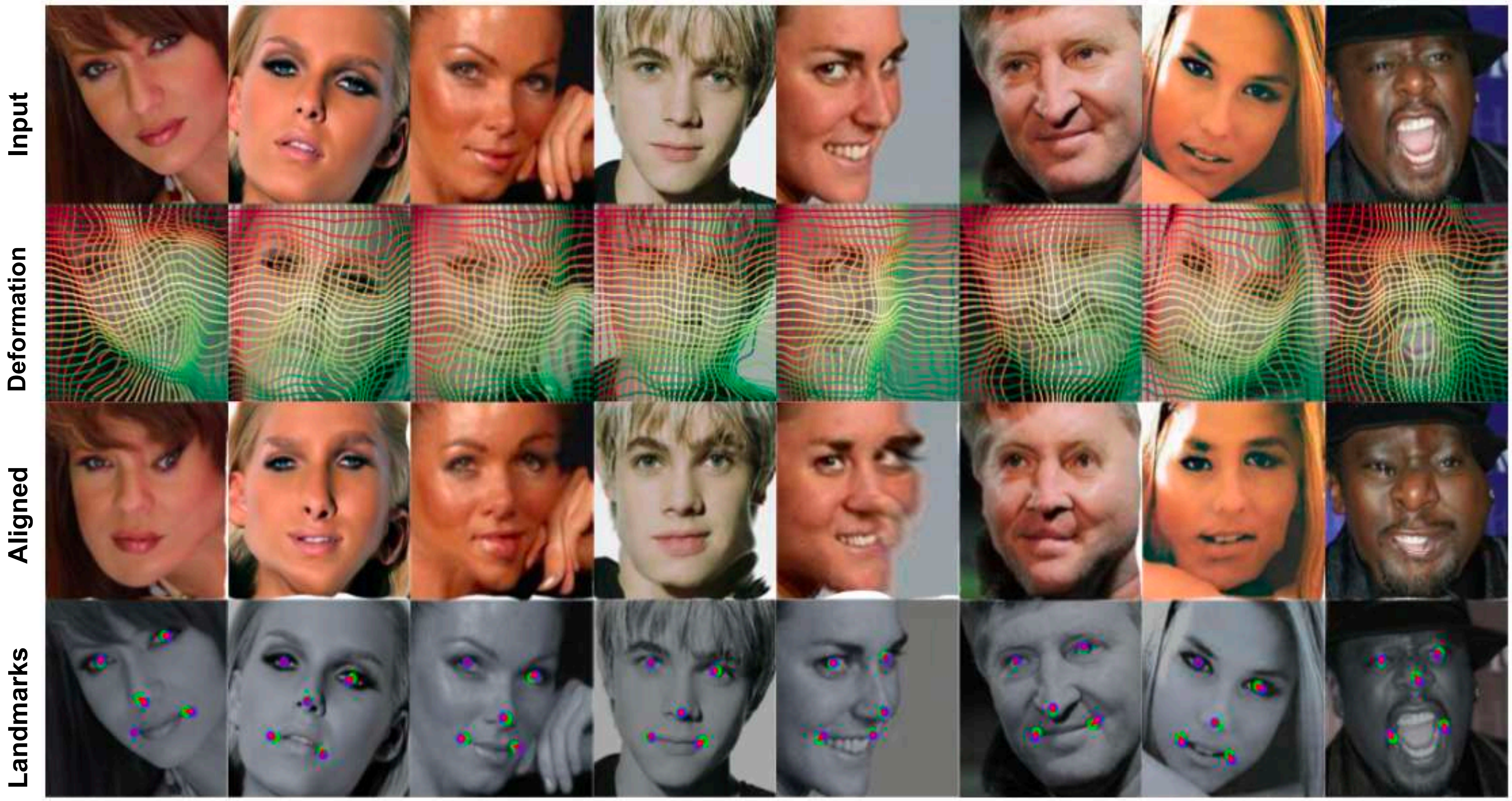
Z. Zhu, M. Saha, A. Guler, D. Samaras, I. Kokkinos,
Deforming Autoencoders: Unsupervised Shape and Appearance Disentangling, ECCV 2018

DAE for MNIST: single-class template



DAE for Faces-in-the-Wild

DAE-based unsupervised face alignment



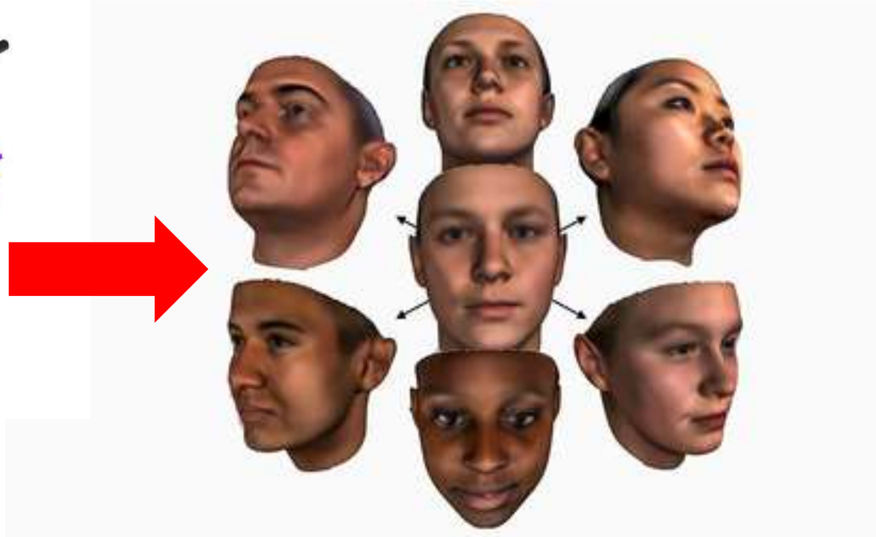
Unsupervised alignment with DAE on MAFL dataset

Goal: learn a 3D model from unstructured image set

Unstructured face dataset

Something deep happens

3D model comes out



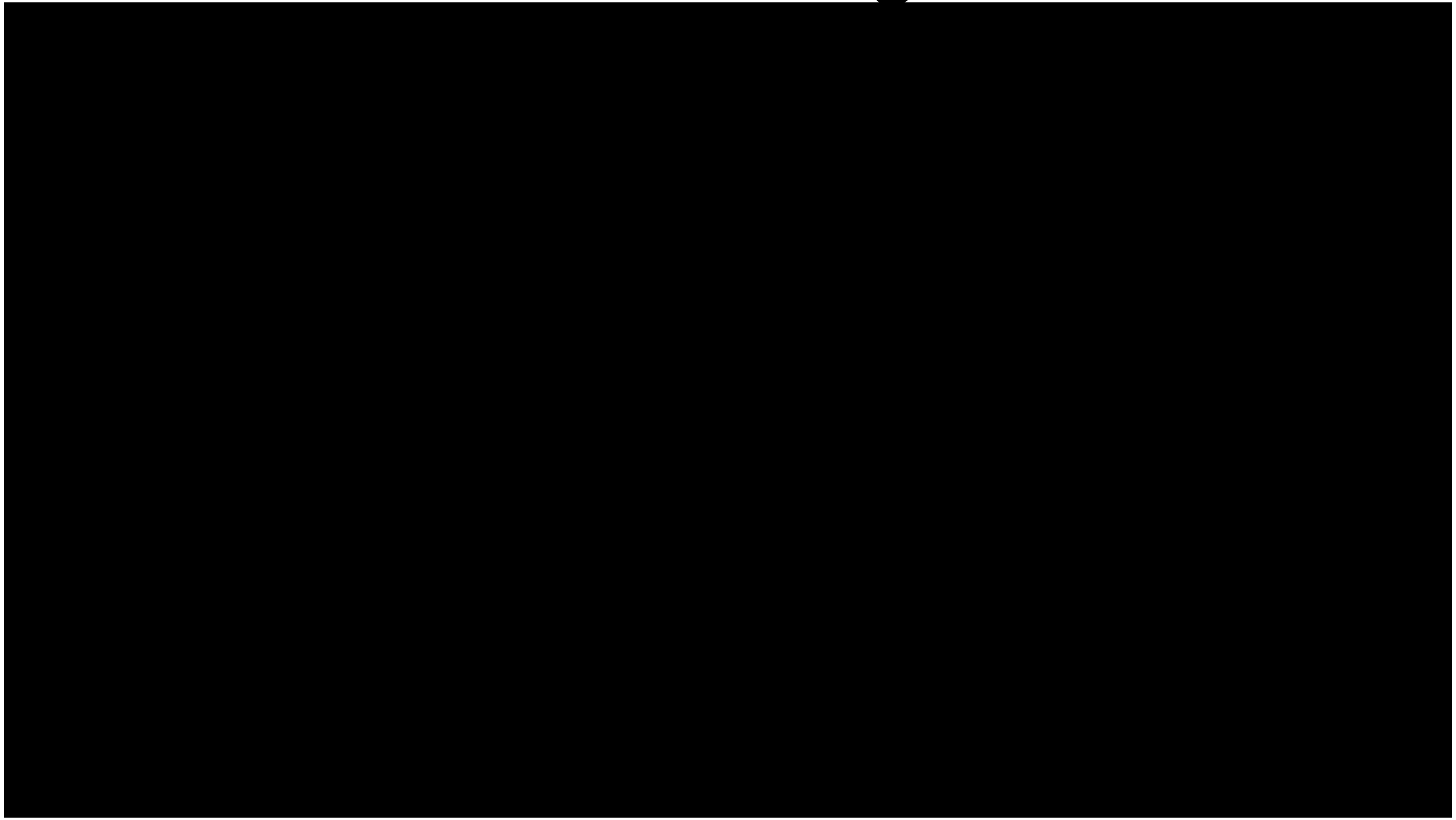
3D Reconstruction: Structure-from-Motion

Assumption: Rigid Scene
Input: Point Correspondences (e.g. through SIFT & Ransac)

Methods: Factorization, Bundle Adjustment

Noah Snavely, Steven M. Seitz, Richard Szeliski. Modeling the World from Internet Photo Collections. IJCV, 2007.
Yasutaka Furukawa and Jean Ponce, Accurate, Dense, and Robust Multi-View Stereopsis, CVPR 2007

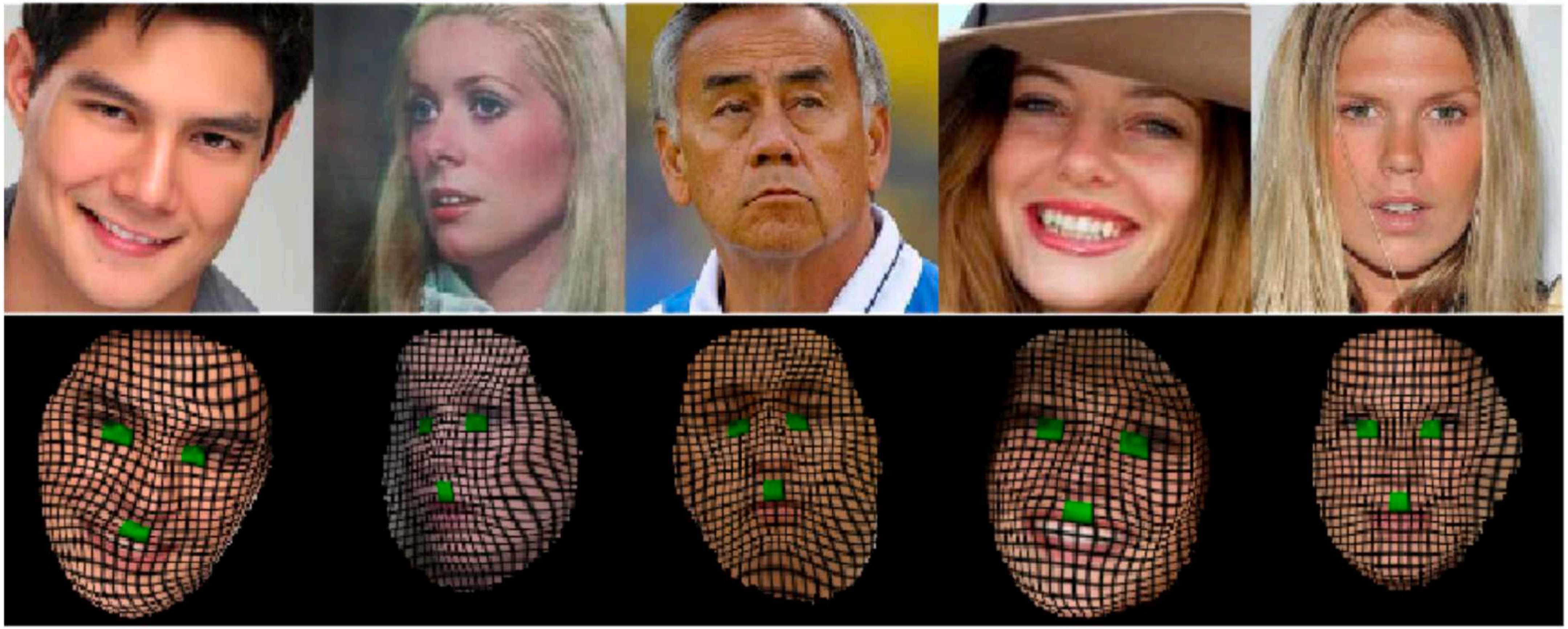
3D Reconstruction: Non-Rigid Structure-from-Motion



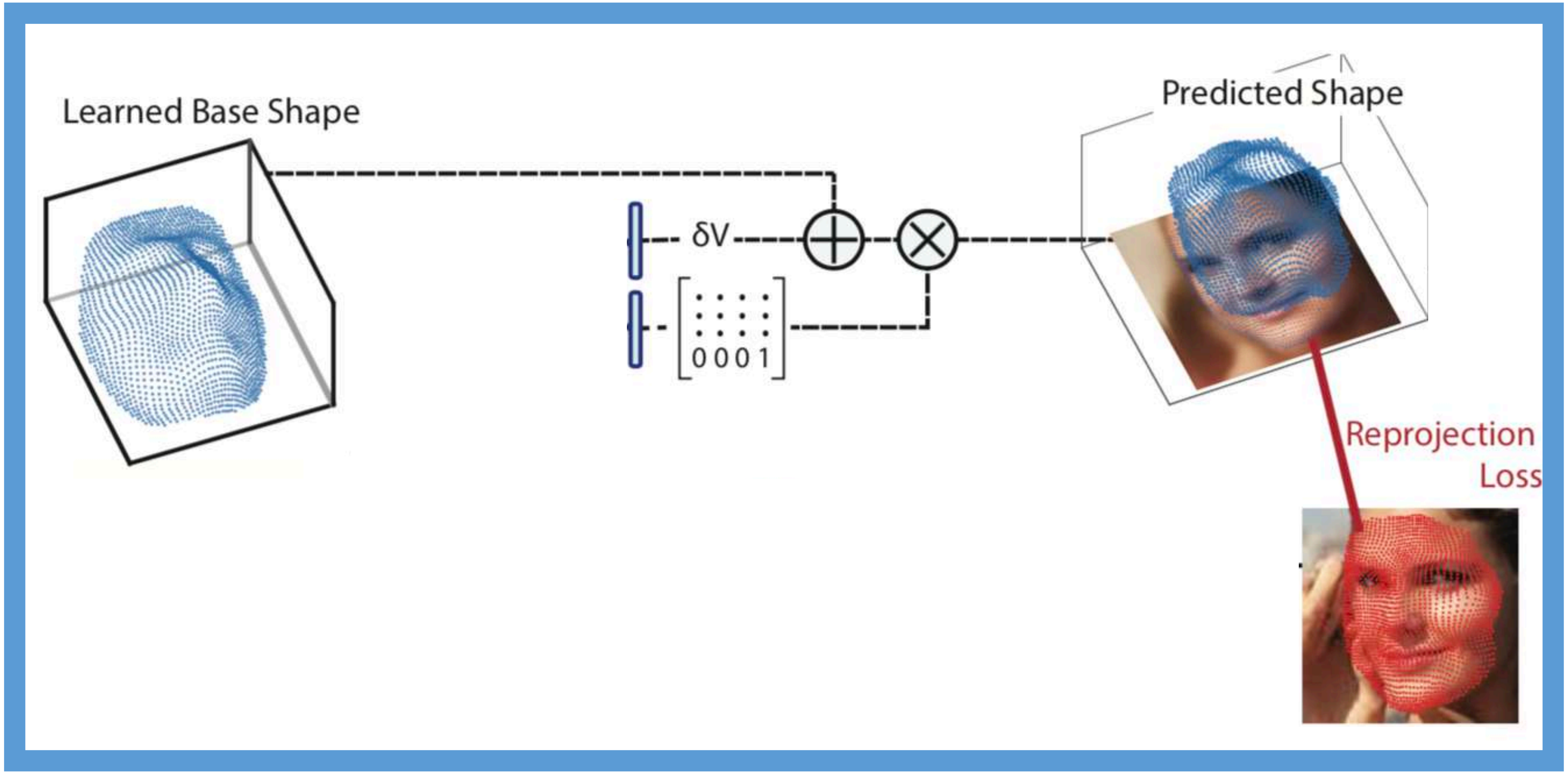
https://www.youtube.com/watch?v=35wCPFyS3QQ

Non-Rigid Structure-From-Motion: Estimating Shape and Motion with Hierarchical Priors, Bregler et al, PAMI 2008
Dense Reconstruction of Non-Rigid Surfaces from Monocular Video, Garg et al, CVPR 2013

DAEs: Turn Images to Corresponding Sets of Points

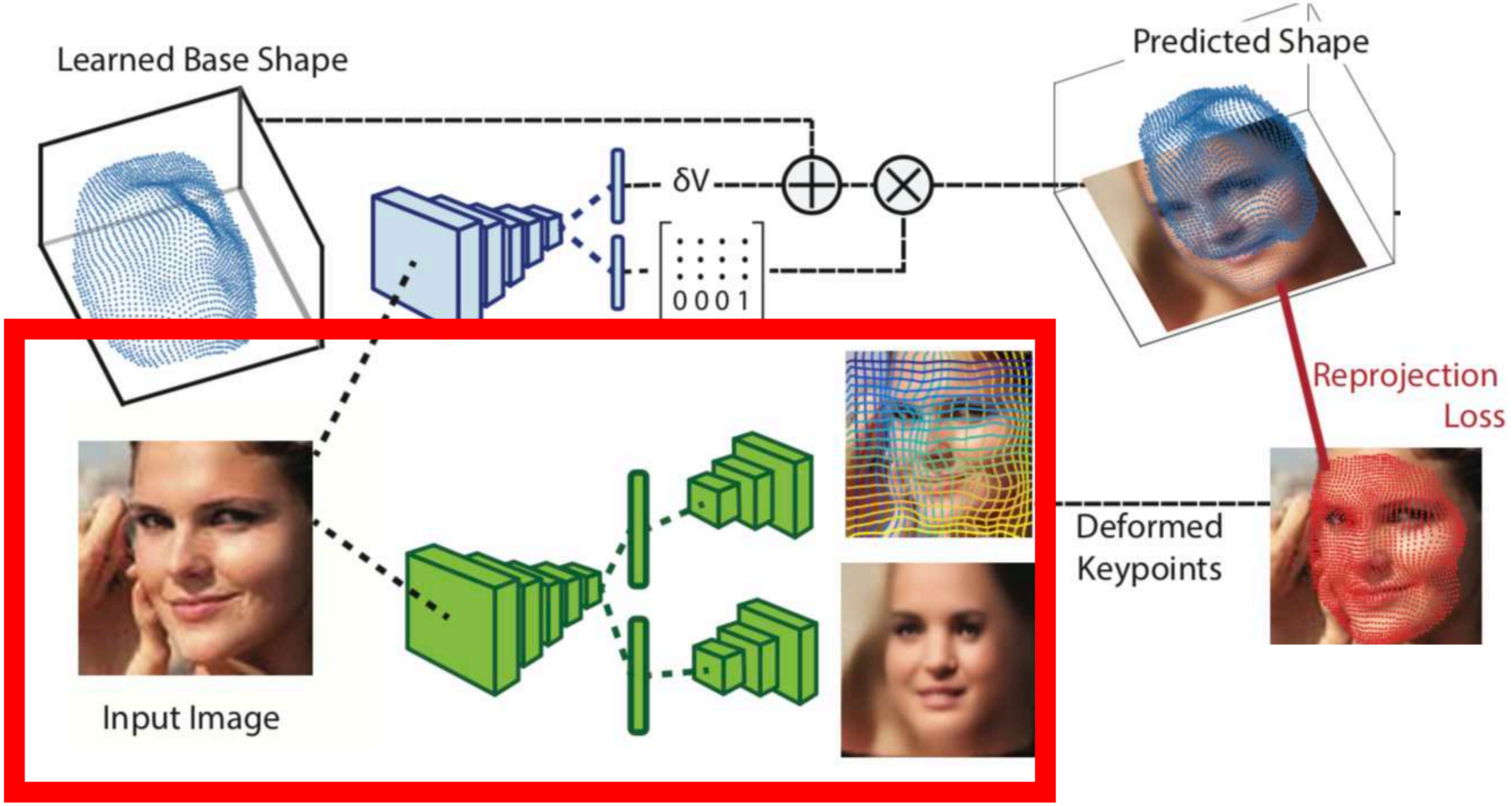


Lifting AutoEncoder: NRSfM with DAEs



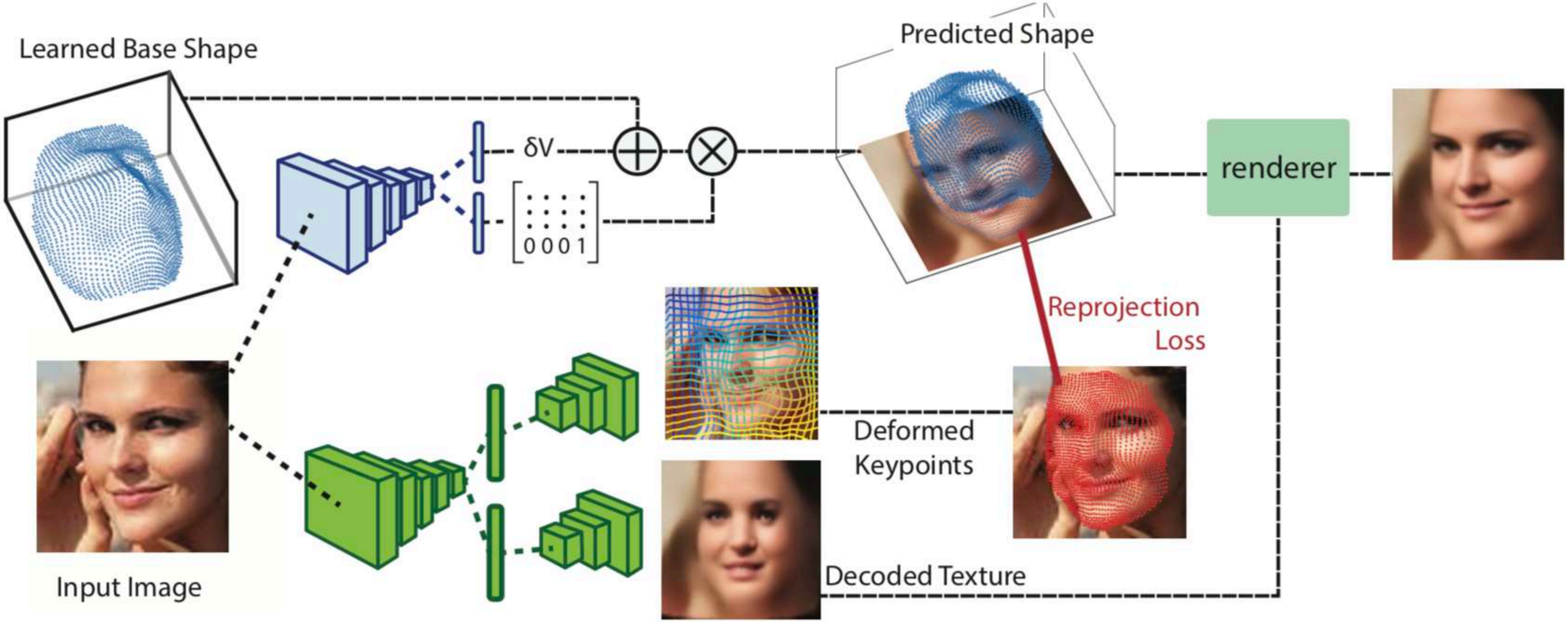
Non-Rigid Structure-from-Motion

Lifting AutoEncoder: NRSfM with DAEs



Deforming AutoEncoder

Lifting Auto-Encoders: end-to-end 3D generative model



Lifting AutoEncoders: Unsupervised Learning of a Fully-Disentangled 3D Morphable M§odel

Controllable image modification using LAEs

Pose modification

Lifting AutoEncoders: Unsupervised Learning of Fully-Disentangled 3D Morphable model

Controllable image modification using LAEs

Pose modification

Expression modification

Lifting AutoEncoders: Unsupervised Learning of Fully-Disentangled 3D Morphable model

Controllable image modification using LAEs

Pose modification

Expression modification

Lifting AutoEncoders: Unsupervised Learning of Fully-Disentangled 3D Morphable model

Controllable image modification using LAEs

Pose modification

Expression modification

Illumination modification

Part 1: Weakly- and semi- supervised learning for 3D

HoloPose: Holistic 3D Human Reconstruction In-the-Wild, A. Guler and I. Kokkinos, CVPR 2019

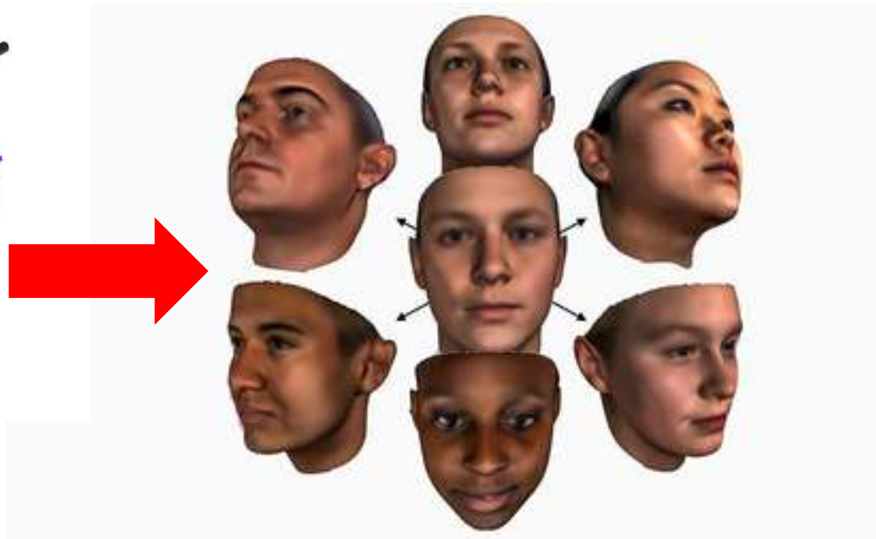
Weakly-Supervised Mesh-Convolutional Hand Reconstruction in the Wild, D. Kulon et al CVPR 2020

Part 2: Fully unsupervised learning for 3D

Unstructured face dataset

deep magic happens

3D model comes out

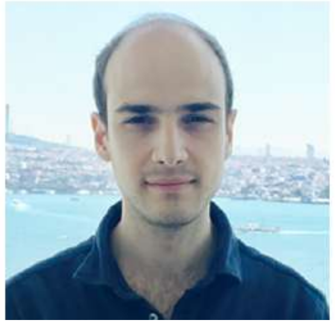


Lifting AutoEncoders: Unsupervised Learning of 3D Morphable Models Using Deep Non-Rigid Structure from Motion, M. Sahasrabudhe, Z. Shu, E. Bartrum, A. Guler, D. Samaras and I. Kokkinos, ICCV GMDL 2019

Thank you!

arielai.com/mesh_hands

Ariel AI



R. A. Guler

G. Papandreou

B. Fulkerson

S. Zafeiriou

E. Schmitt

H. Wang

D. Kulon

P. Koutras

E. Skordos

S. Galanakis

A. Kakolyris

D. Stoddard

H. Tam

A. Lazarou

M. Bronstein
Imperial College

Natalia Neverova
FAIR

Z. Shu
Stony Brook

M. Sahasrabudhe
INRIA

E. Bartrum
UCL

N. Paragios
INRIA

D. Samaras
Stony Brook