Virtual Humans – Winter 23/24

Lecture 6_2 – ICP: Fitting SMPL to Images with Learning

Prof. Dr.-Ing. Gerard Pons-Moll

University of Tübingen / MPI-Informatics

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

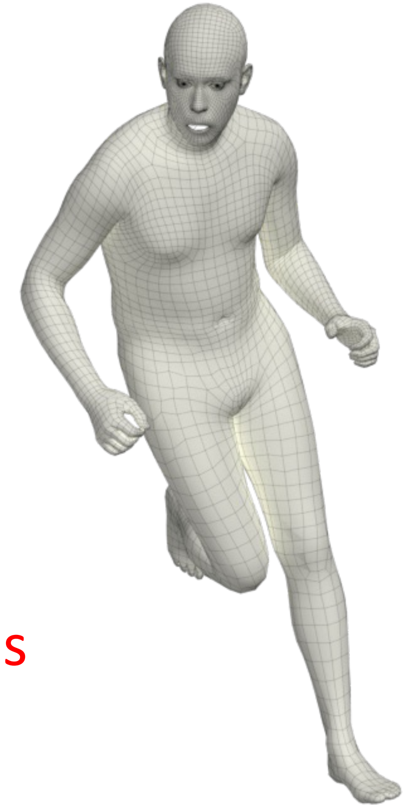
Goal: Estimate SMPL from a single image

Estimate 3D shape and pose

"See" the person in 3D

Problem with optimization based fitting

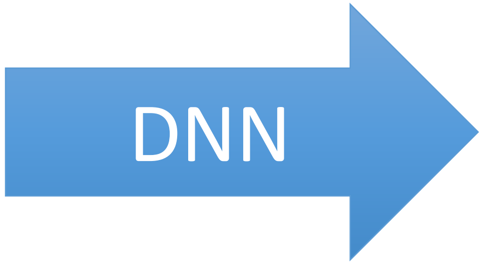
Fitting

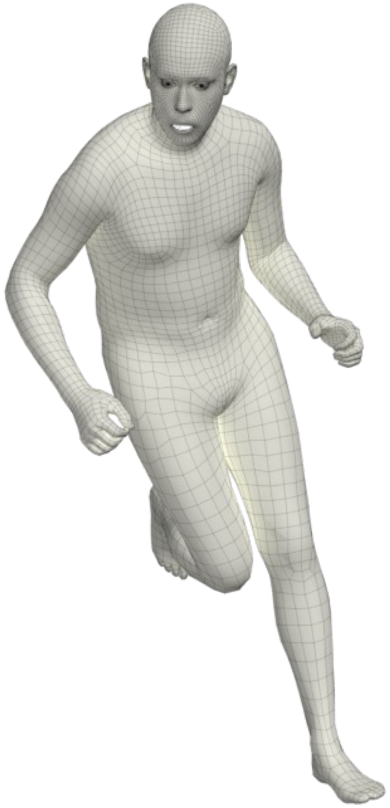


- Requires pre-defined features
- Slow
- Local minima

Can we use learning to get better SMPL?

Learn a mapping directly from image pixels to SMPL parameters using a DNN.

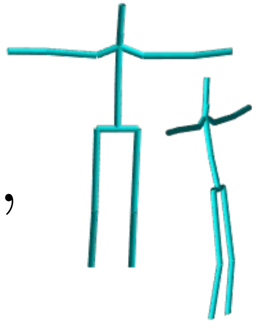




DNN = Deep Neural Network

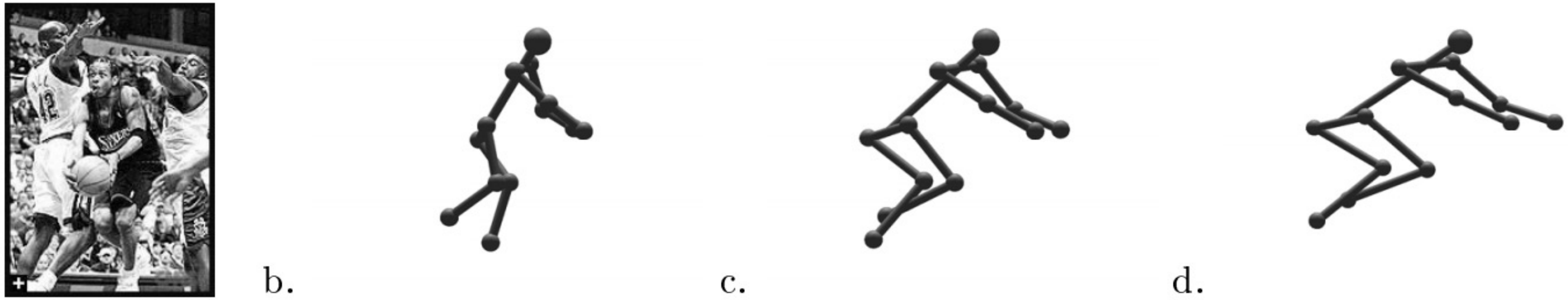
Challenges

- Lack of real paired 2D-to-3D data

{  ,  }

{  , 3D?? }

- Depth ambiguity

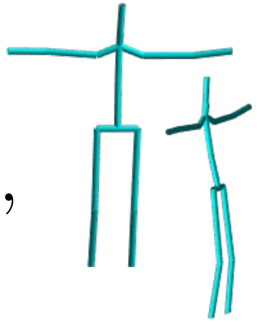


a. b. c. d.

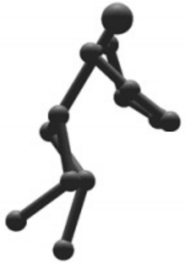
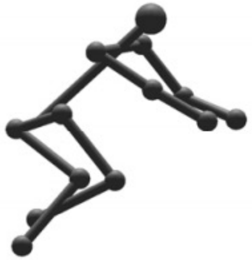
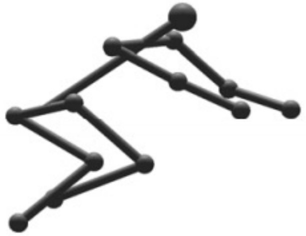
[CJ Taylor CVPR 2000]

Ideas…?

- Can we train a neural network with only 2D supervision?

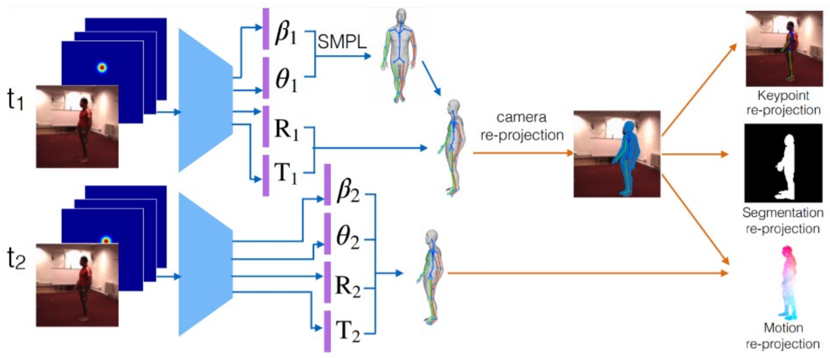
{  ,  } {  , 3D?? }

- Can we learn prior using unpaired 2D-3D data?

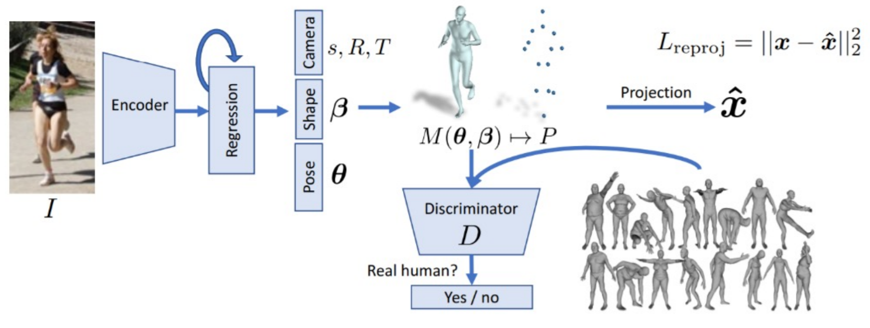
a.  b.  c.  d. 

[CJ Taylor CVPR 2000]

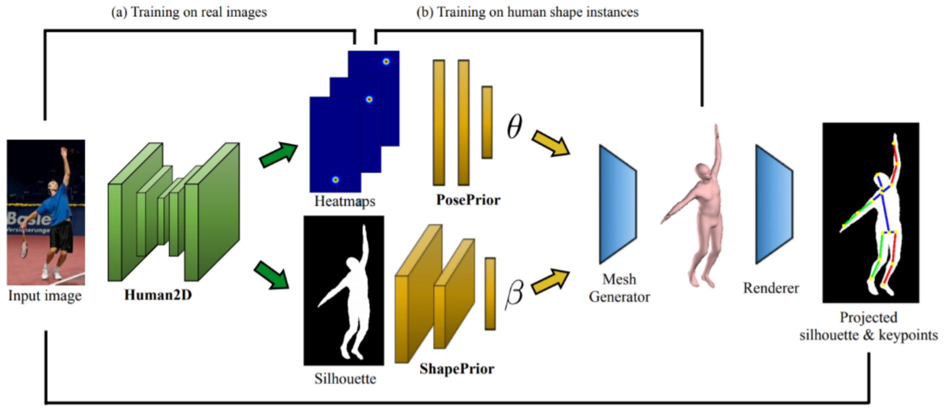
Self-supervised hybrid approaches



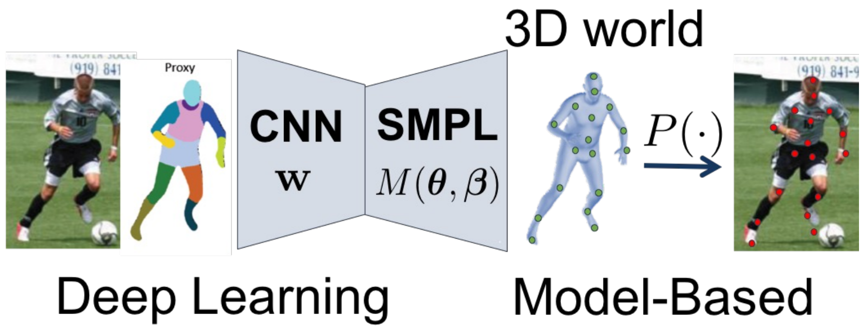
(Tung et al. 2017)



(Kanazawa et al. 2018)



(Pavlakos et al. 2018)



Omran et al. 2018 (3dV Best student paper award)



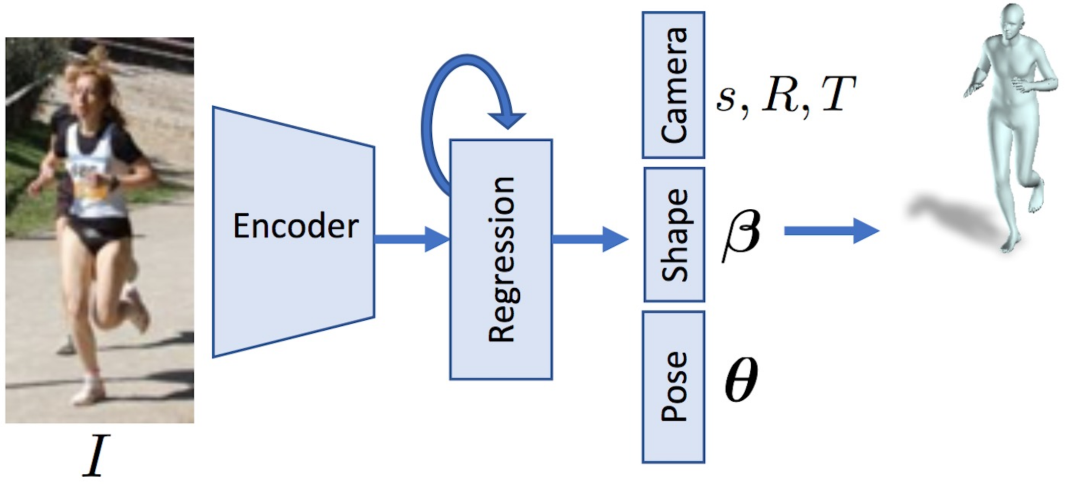

End-to-end Recovery of Human Shape and Pose (HMR)

A. Kanazawa, M. J. Black, D. W. Jacobs, J. Malik

CVPR'18

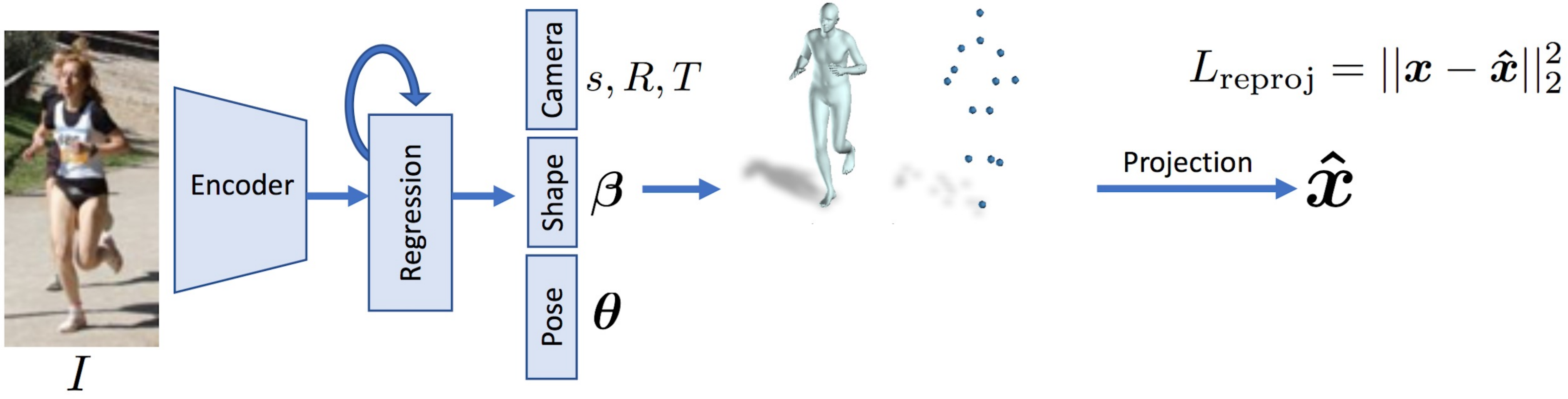
Some of the following slides are adapted from slides provided by Kanazawa et al.

Goal: Predict 3D SMPL without paired data



Question: How to learn a deep neural network to directly regress SMPL parameters without any paired 3D supervision?

Train a neural network with 2D supervision?



Produces monsters!

Can we regularise the predicted SMPL?

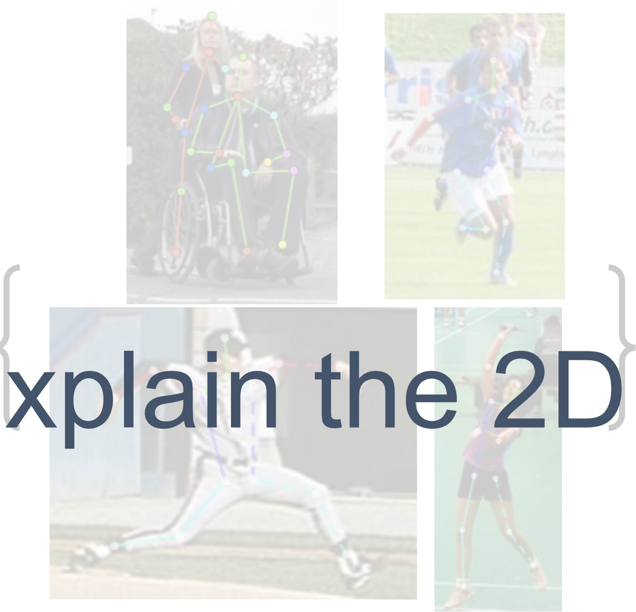
Large 2D and 3D datasets exist

2D Labeled images
[LSP, MPII,COCO,…]

3D Scans/Motion Capture
[CMU Mocap, CAESER, JointLimits..]

Can we regularise the predicted SMPL?

Leverage **unpaired** data



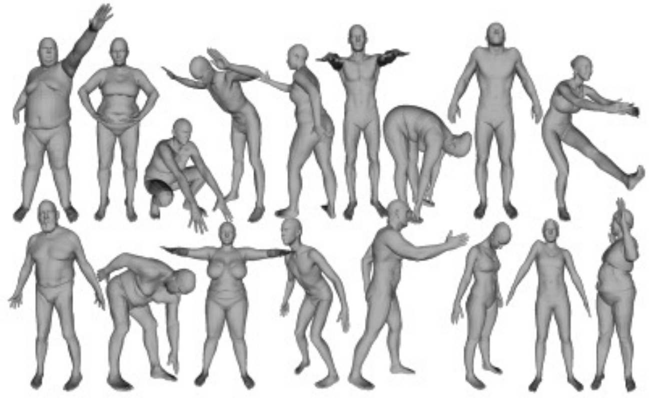
2D Labeled images
[LSP, MPII,COCO,…]

3D Scans/Motion Capture
[CMU Mocap, CAESER, JointLimits..]

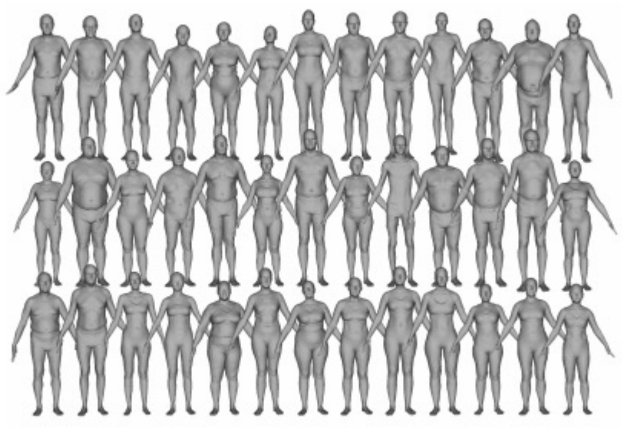
Can we regularise the predicted SMPL?

We have used pose and shape prior before during optimization!

GMM based prior in SMPLify

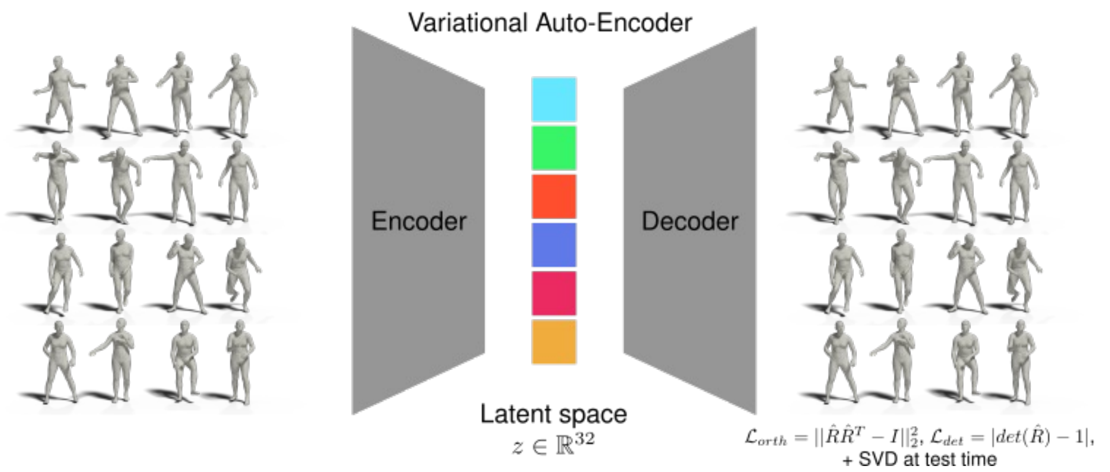


$E_\theta(\vec{\theta})$ Pose Prior



$E_\beta(\vec{\beta})$ Shape Prior

VAE based prior in SMPLify

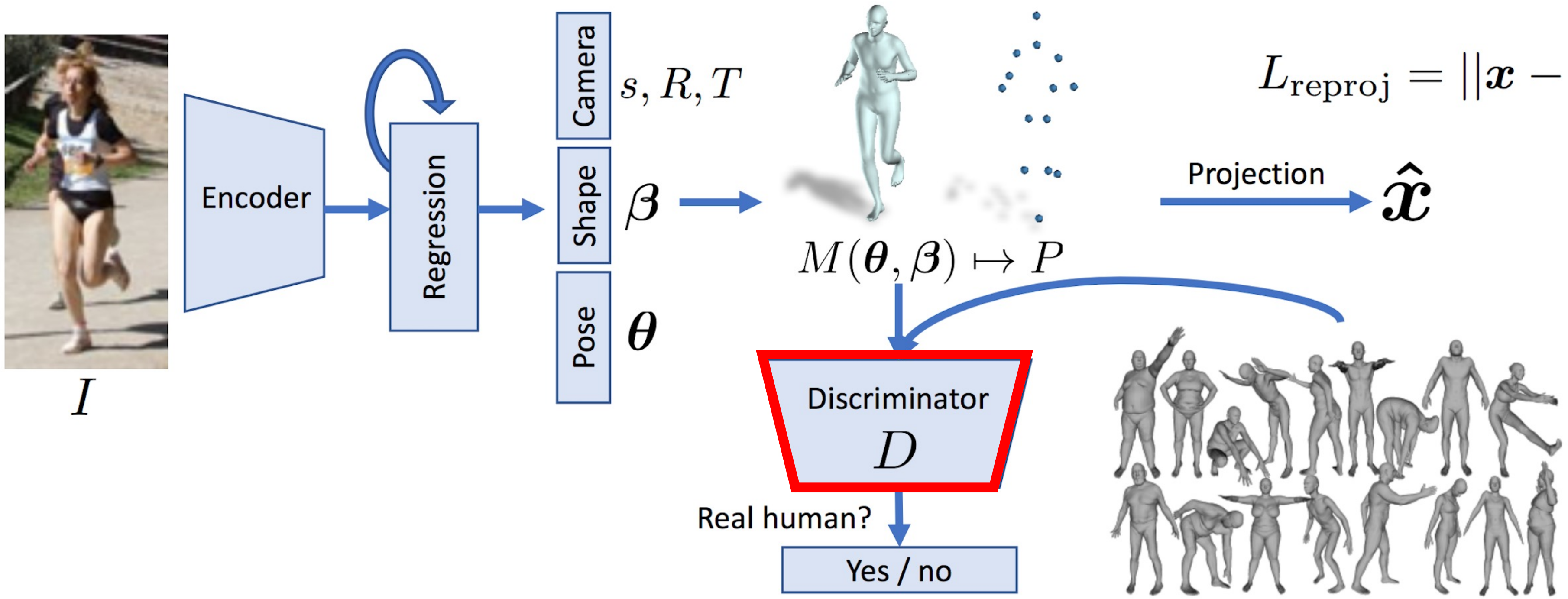


What prior can be used?

- A prior models the natural distribution and estimates the likelihood that a sample belongs to the distribution.

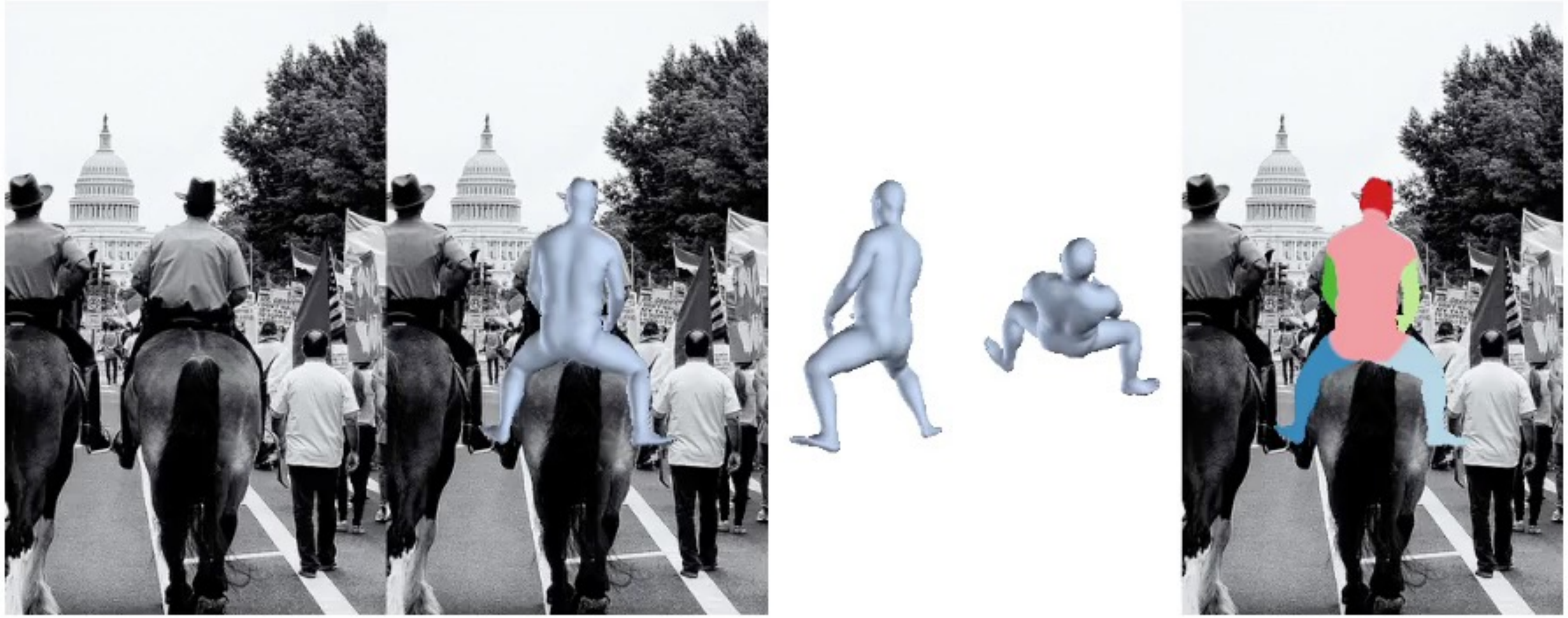
- Is there another very popular way to capture data distribution?
- Yes, GAN

Direct regression from pixels?



The adversary (D) knows about body shape and pose.

Results from HMR

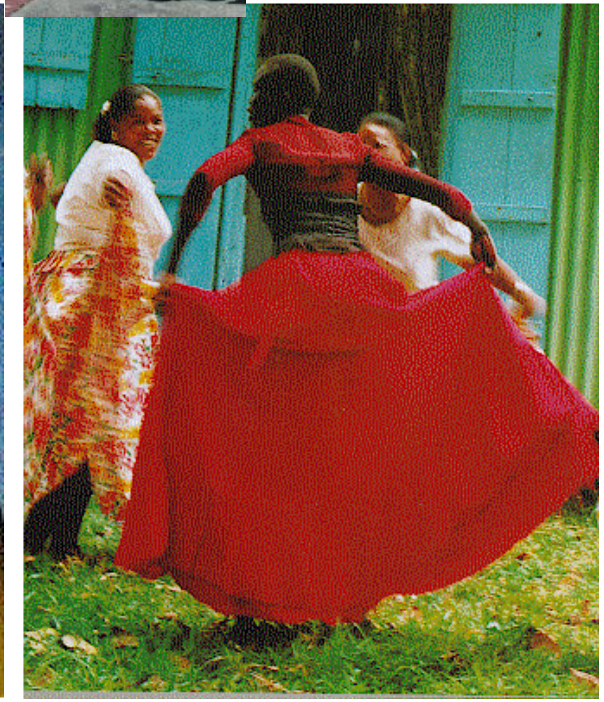


Input Reconstruction Part segmentation

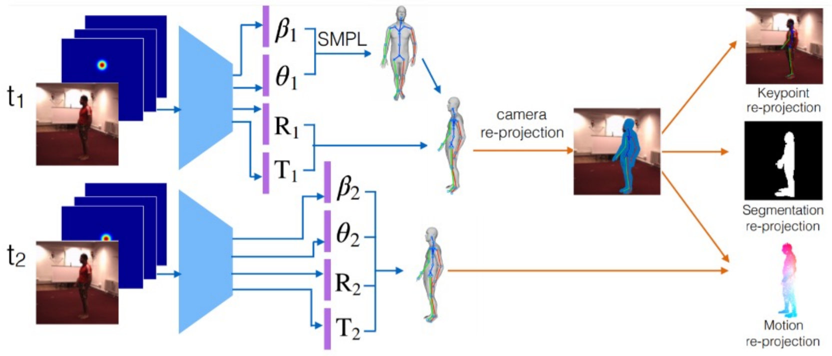
Remaining problem:
Large variability in **appearance**



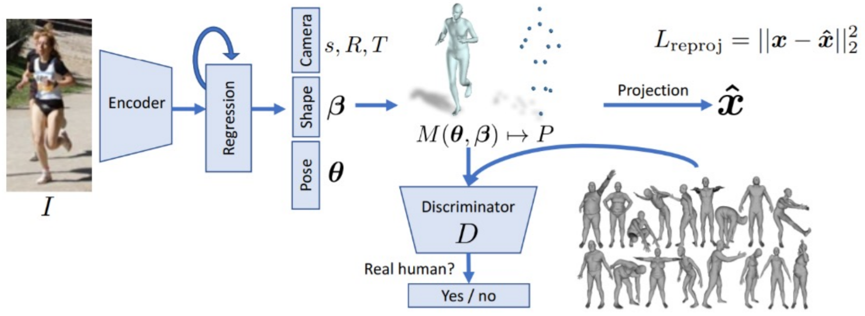


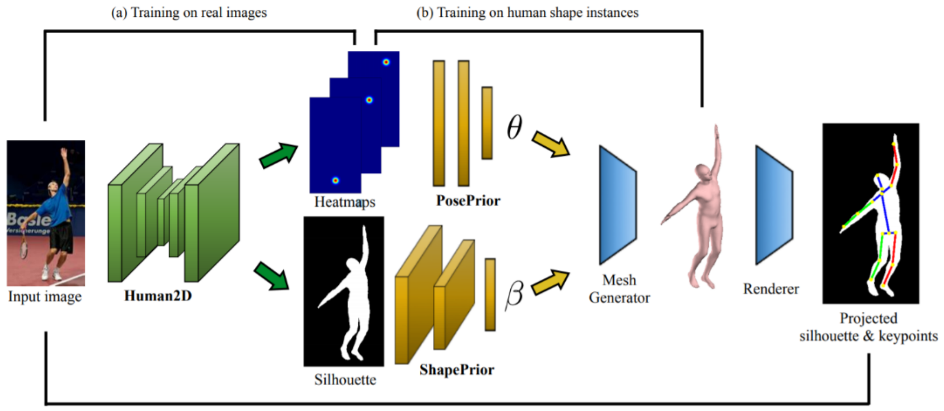
Self-supervised hybrid approaches



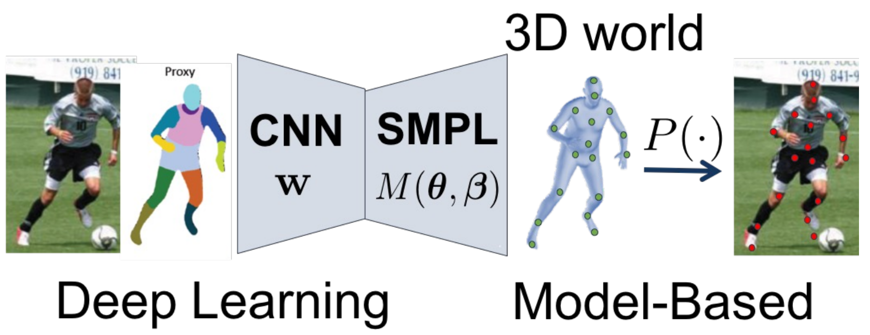
(Tung et al. 2017)



(Kanazawa et al. 2018)



(Pavlakos et al. 2018)



Omran et al. 2018 (3dV Best student paper award)





Neural Body Fitting (NBF):

Unifying Deep Learning and Model-Based Human Pose and Shape Estimation

M. Omran, C. Lassner, G. Pons-Moll, P.V. Gehler and B. Schiele

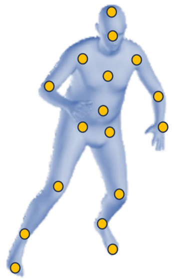
3DV’19 (Best student paper award)

Some of the following slides are adapted from slides provided by Omran et al.

Model-Based Approaches

$$\arg\min_{\boldsymbol{\theta},\boldsymbol{\beta}} \mathrm{dist}\left(\hat{\mathbf{z}}\left(M(\boldsymbol{\theta},\boldsymbol{\beta})\right),\mathbf{z}\right)$$

3D world



$\xrightarrow{P(\cdot)}$

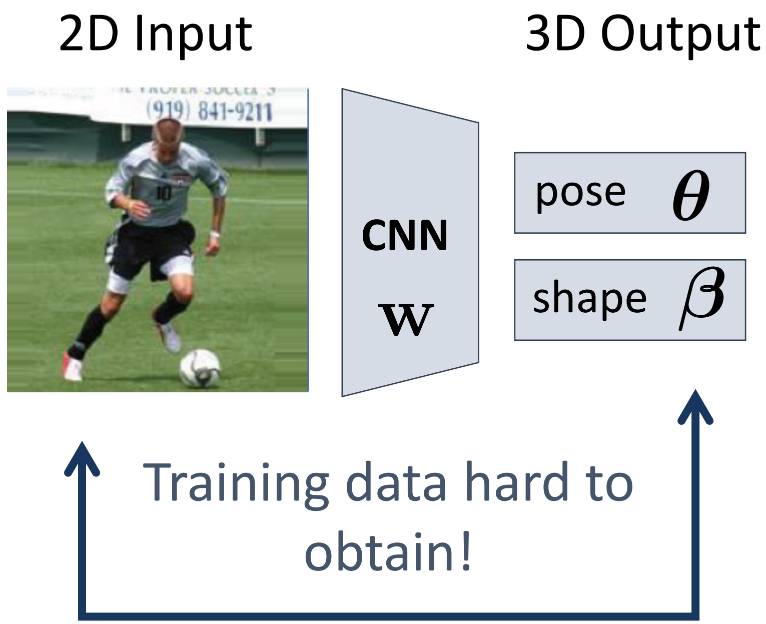
2D keypoints $\mathbf{z}$

$\hat{\mathbf{z}}\left(M(\boldsymbol{\theta},\boldsymbol{\beta})\right)$

Bogo et al. ‘16
Lassner et al. ’17

Optimization can be **slow and complicated**
Optimization requires **careful initialization**

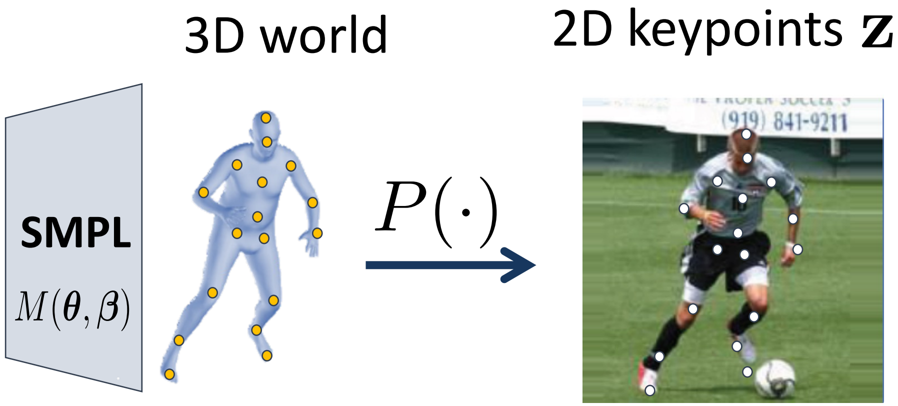
Learning-Based Approaches



Also: no feedback between estimates and observations

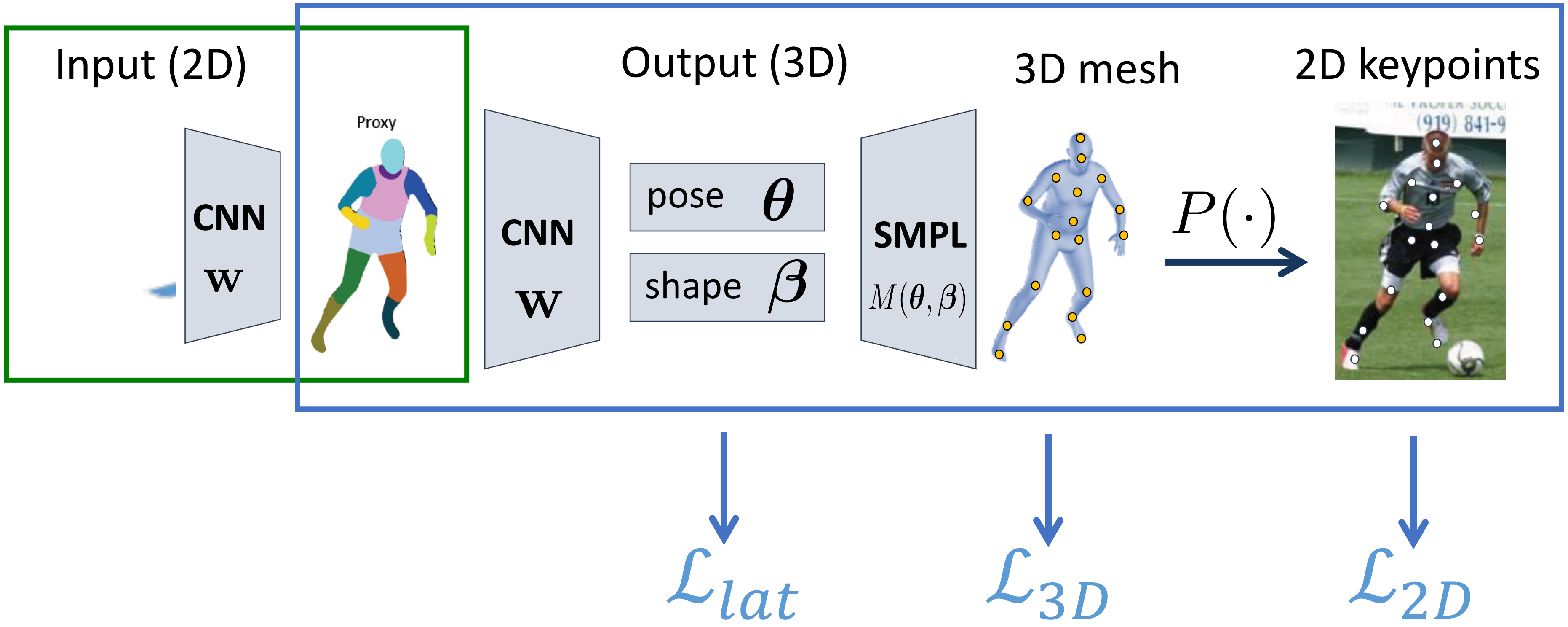
Our Hybrid Approach

Combines aspects of model- and learning-based approaches



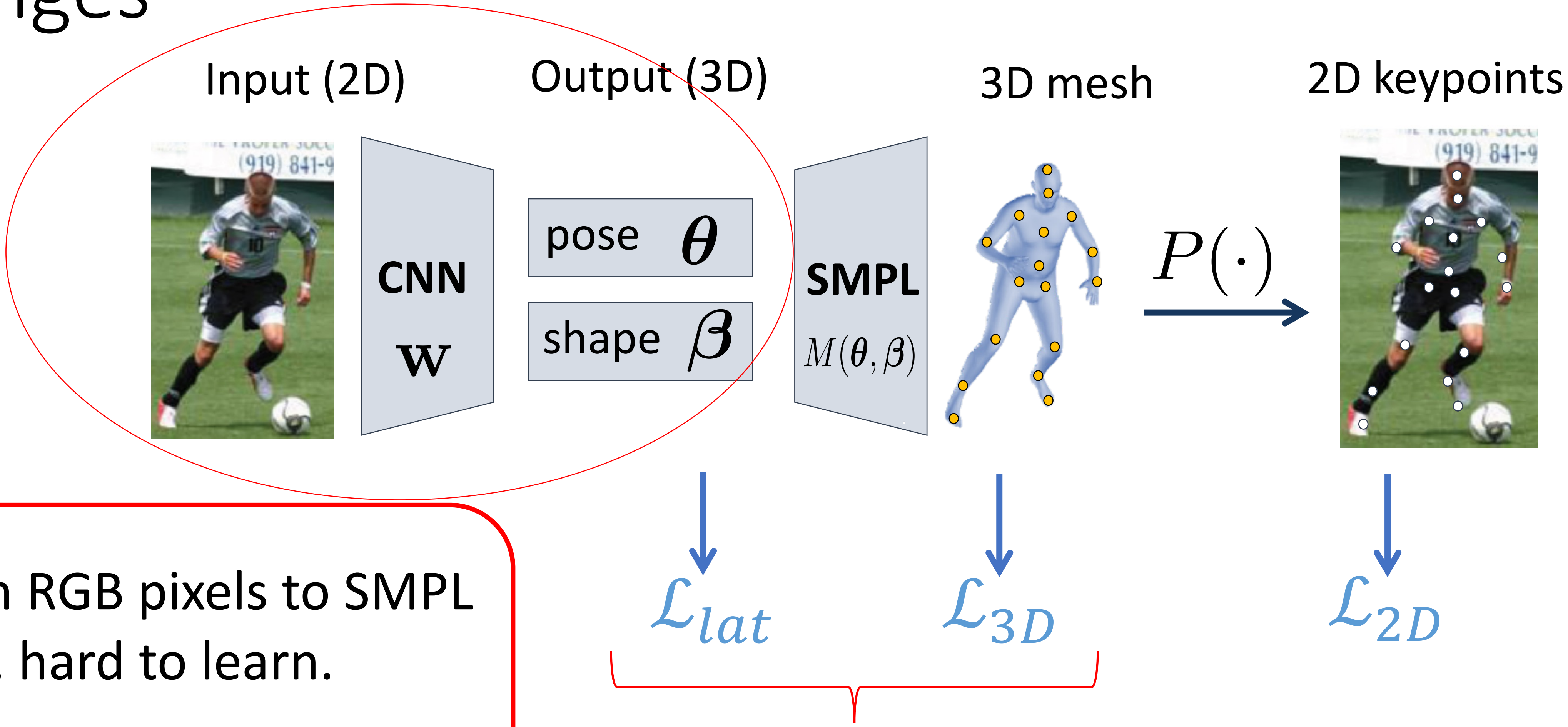
$\mathcal{L}_{lat}$ $\mathcal{L}_{3D}$ $\mathcal{L}_{2D}$

Key research questions



1) Use intermediate 2D representation?
2) Amount of 2D vs 3D supervision?

Challenges



Mapping from RGB pixels to SMPL params. hard to learn.

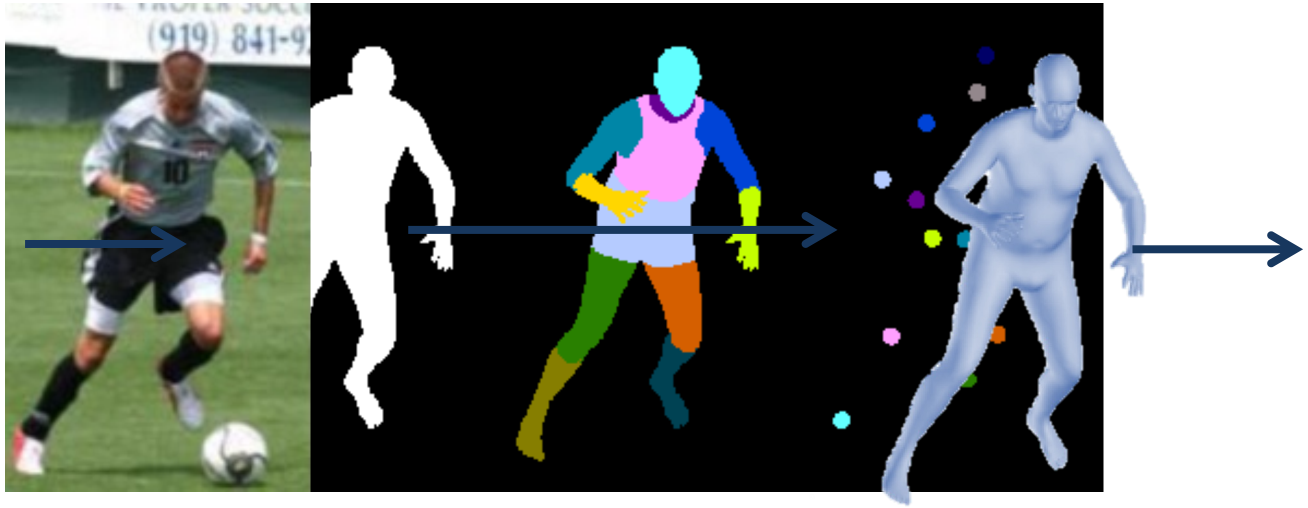
Too much variability in input.

3D data is scarce.

Input Representation

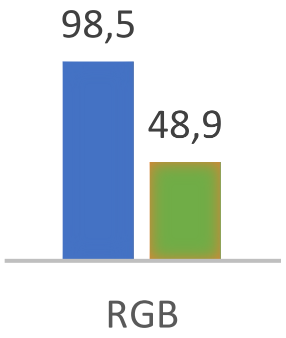
Mapping directly from 2D image to 3D shape and pose is challenging.

Input (2D) Proxy representation Output (3D)



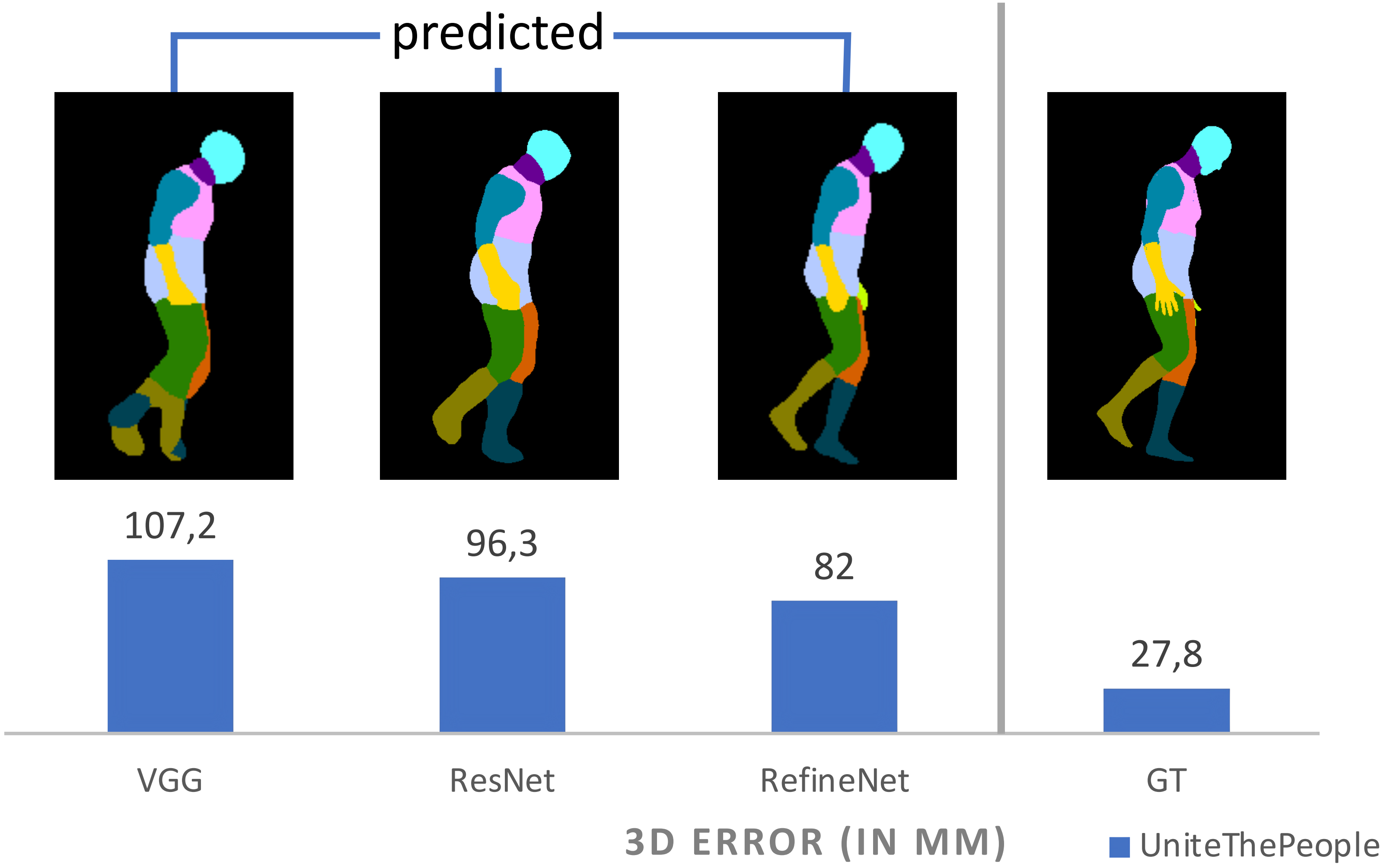
Would an intermediate representation help?
If yes, which?

Input Representation



Lets work with Part Segmentation

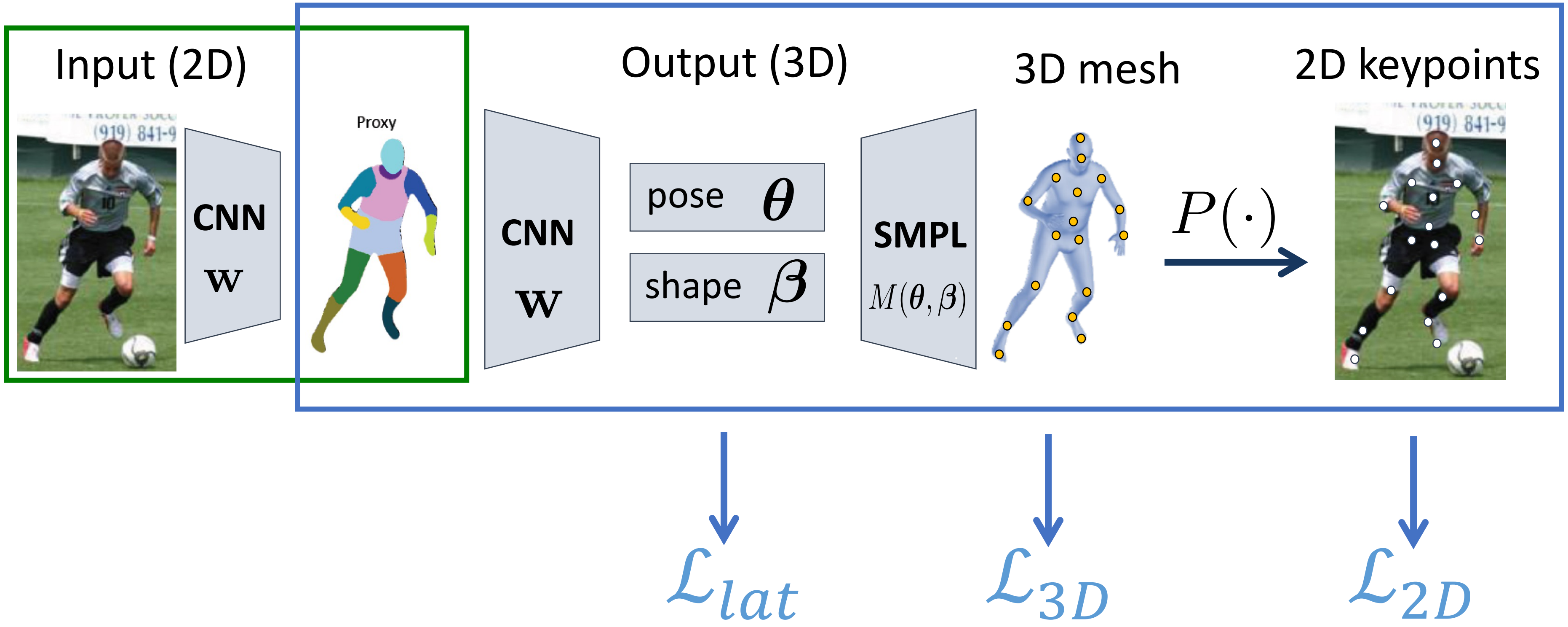
How important is segmentation quality?



Segmentation corelated with Pose Accuracy

- Use part segmentation as intermediate representation.
- Good segmentation is crucial for good 3D shape and pose estimate.

Our Hybrid Approach



1) Use intermediate 2D representation?
2) Amount of 2D vs 3D supervision?

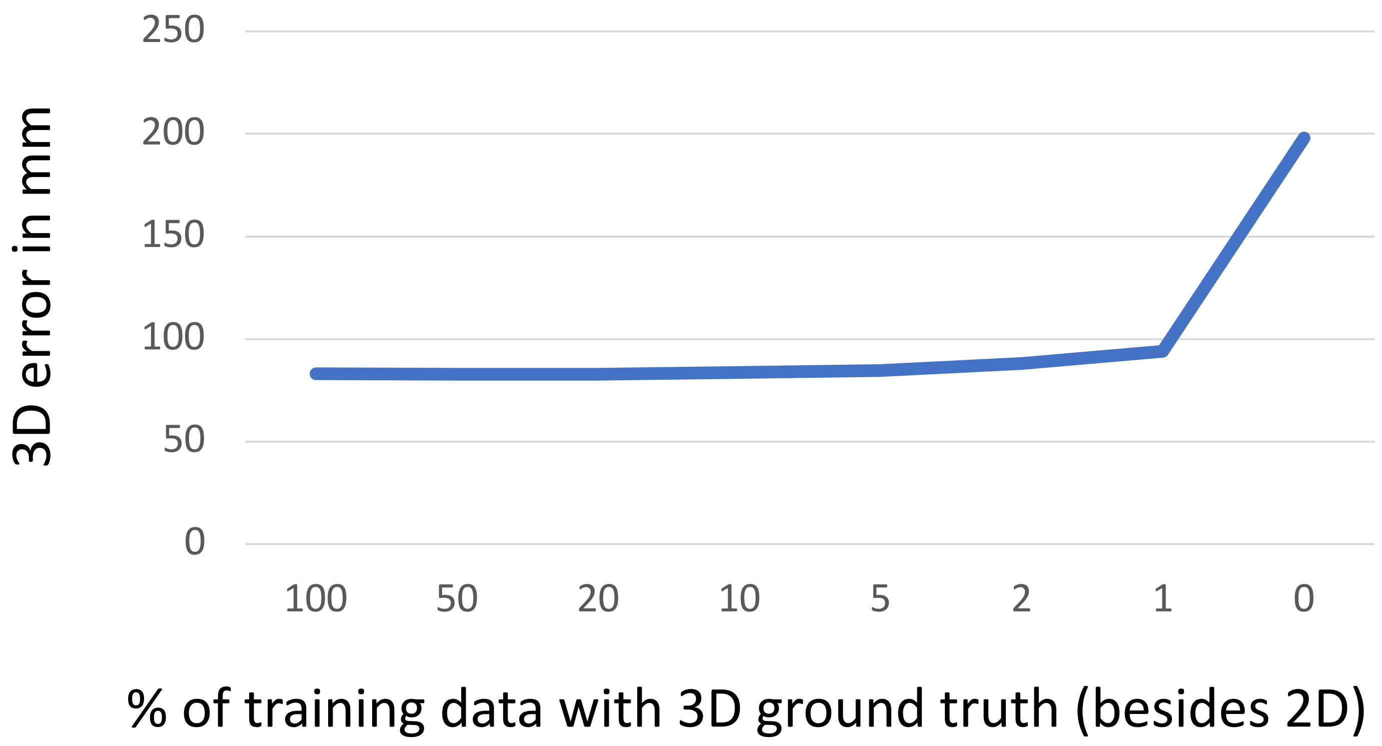
Which Type of Supervision

Loss	Errors		
	3D joints (in mm)	**2D joints (PCKh)**	**joint rotation (in quat.)**
$\mathcal{L}_{2D}$	198.0	94.0	1.971
$\mathcal{L}_{3D}$	83.7	93.5	1.962
$\mathcal{L}_{lat}$	83.7	93.1	0.278
$\mathcal{L}_{lat} + \mathcal{L}_{3D} + \mathcal{L}_{2D}$	82.0	93.5	0.279

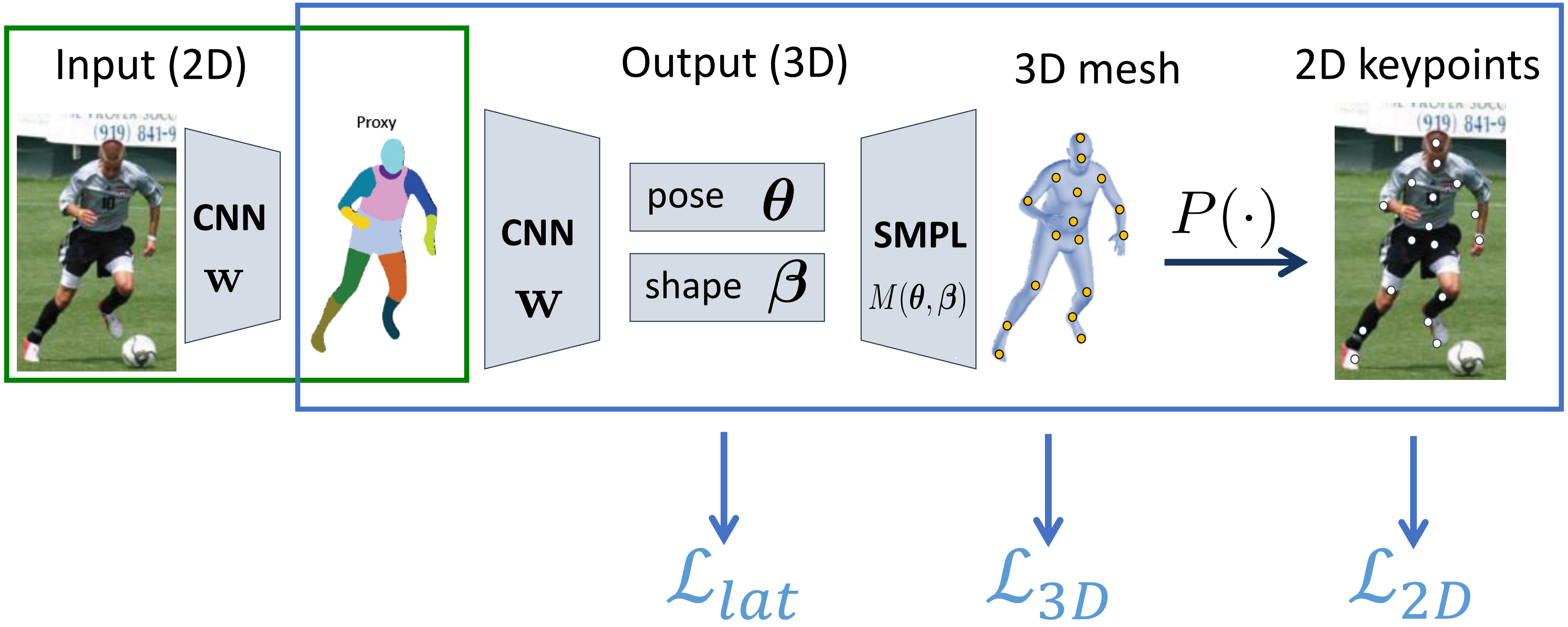
- Supervising with SMPL parameters:
  -> better joint localization (in 2D and 3D) + joint rotations

How Much 3D Supervision?

Experiment: given training data with 2D ground truth (keypoints)
vary size of subset that also has 3D ground truth (shape/pose)



Key messages

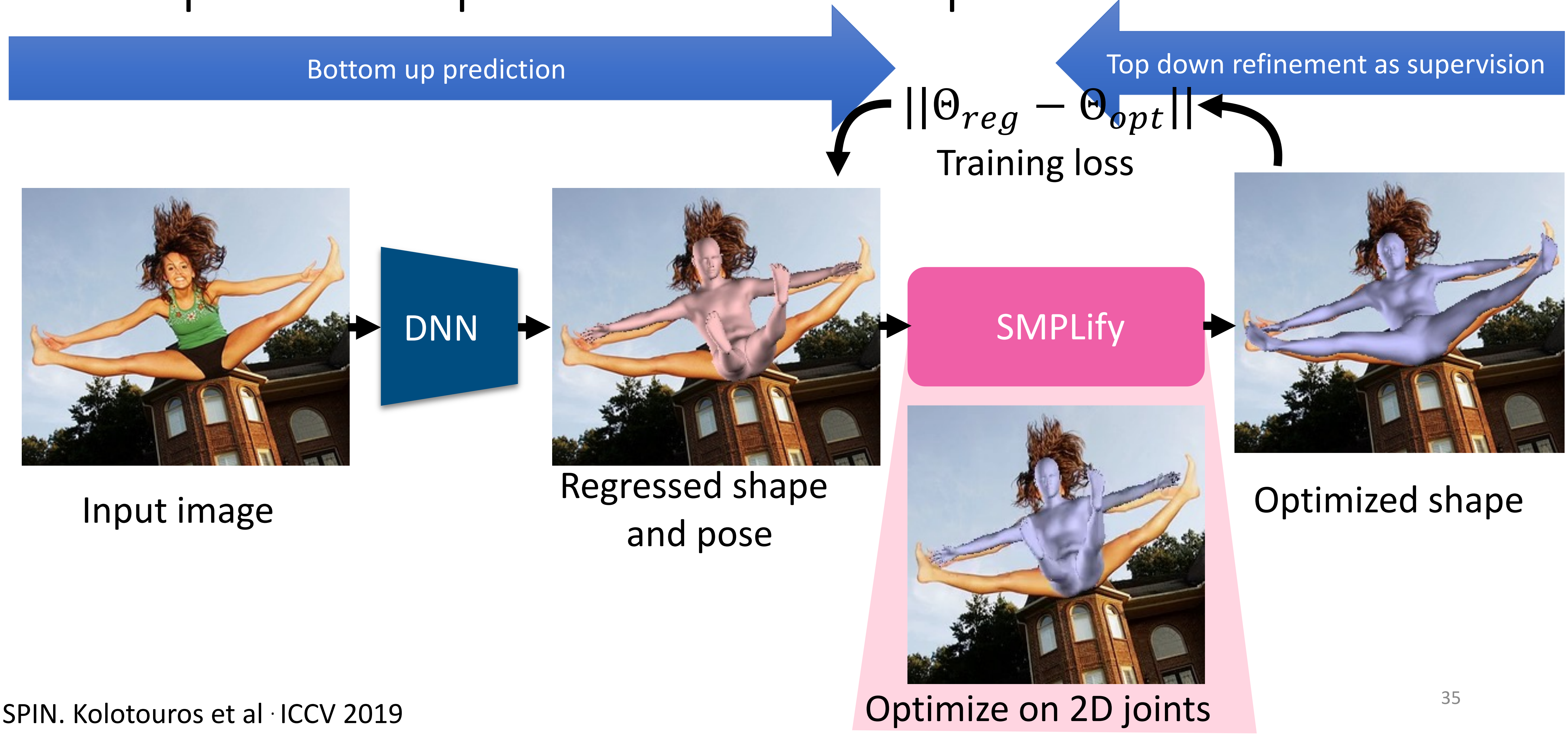


1) Use intermediate 2D representation.
2) Small amount of 3D data is enough.

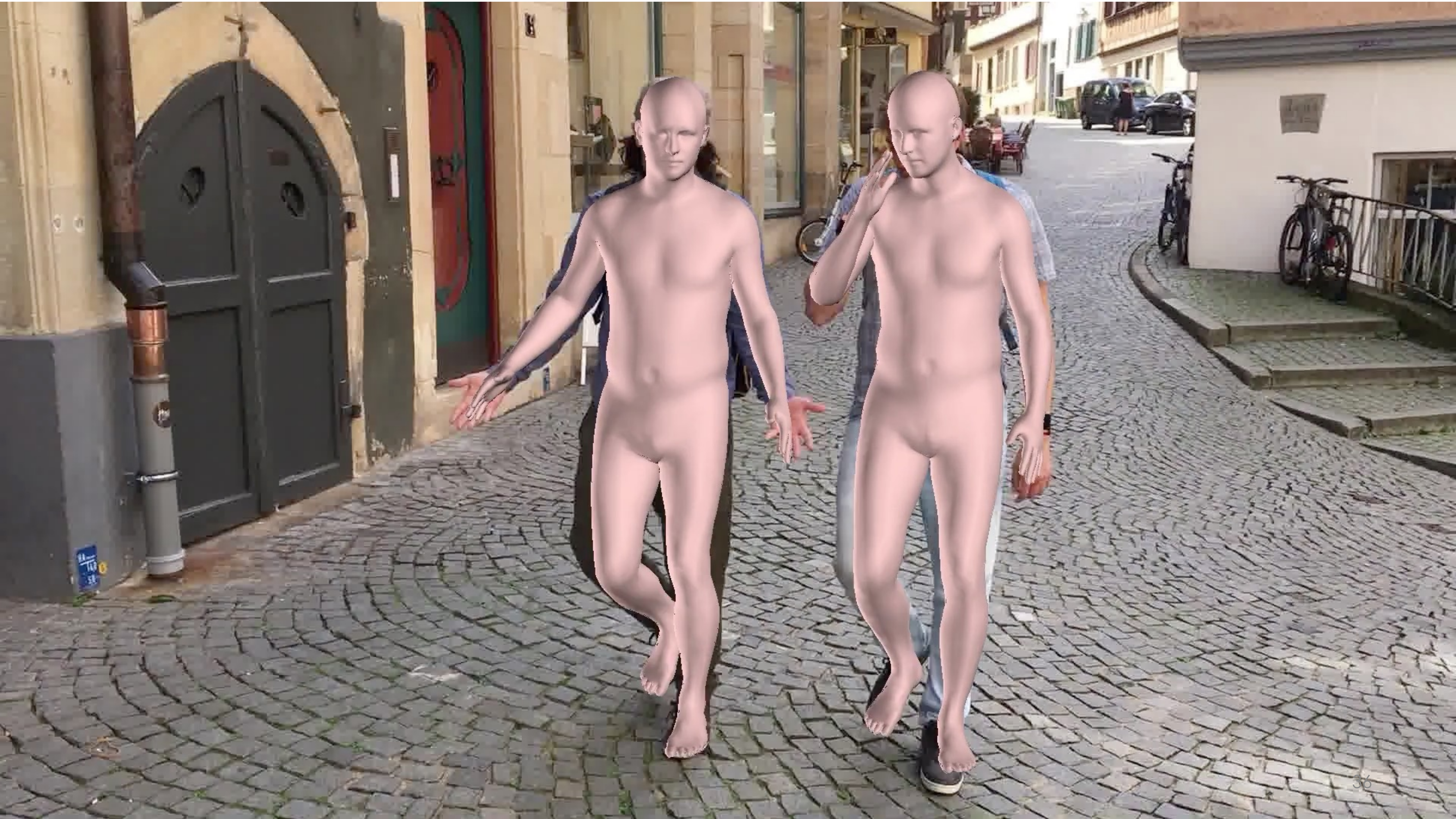
Code is available at:
https://github.com/mohomran/neural_body_fitting

Qualitative Results

Top down optimization as supervision!



SPIN. Kolotouros et al · ICCV 2019



Compare optimization and learning based fitting

Optimization (eg. SMPLify)

✓Better **accuracy,** if initialised well.

✓**Feedback loop**

- Initialization is required

Learning based (eg. HMR)

✓**Automatic**

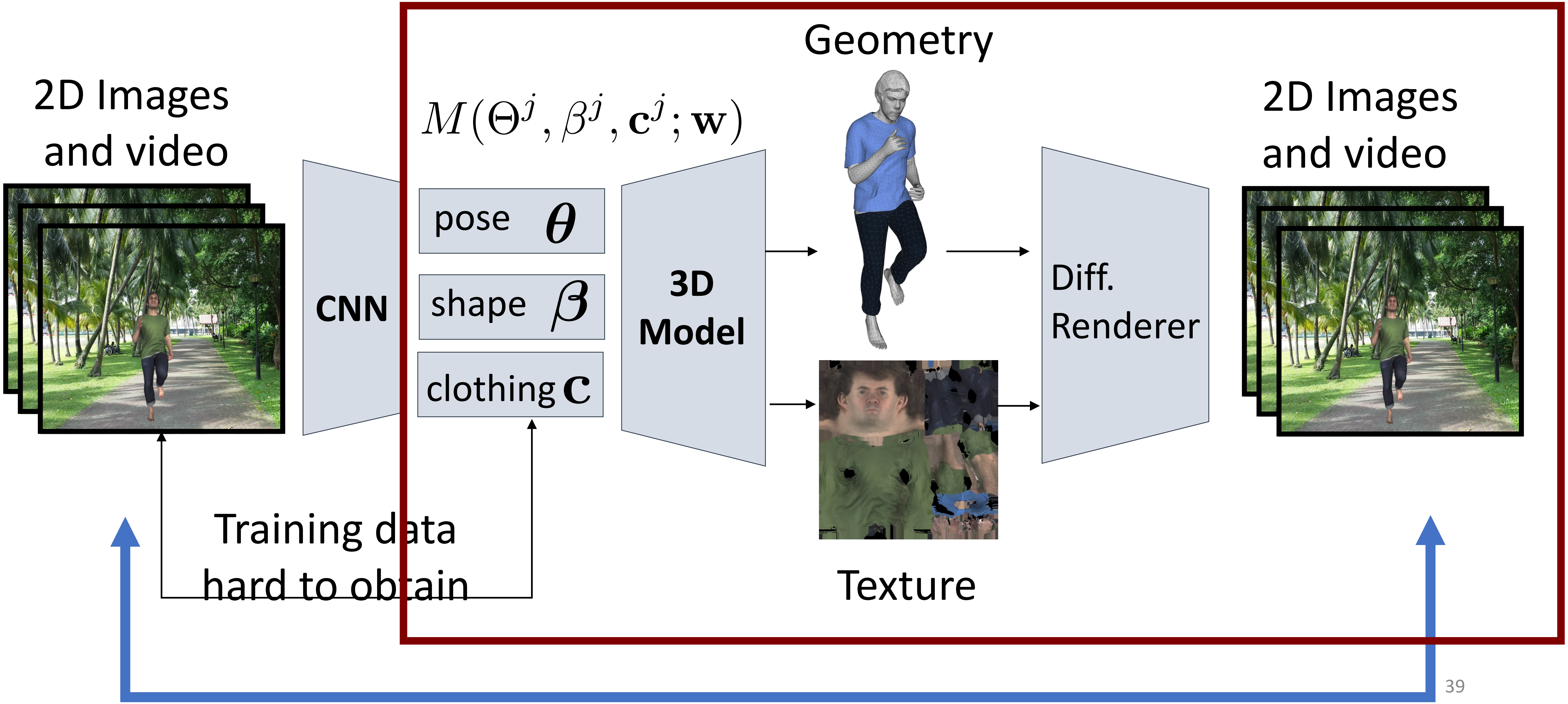
✓ Leverages data prior

–Lower **accuracy.**

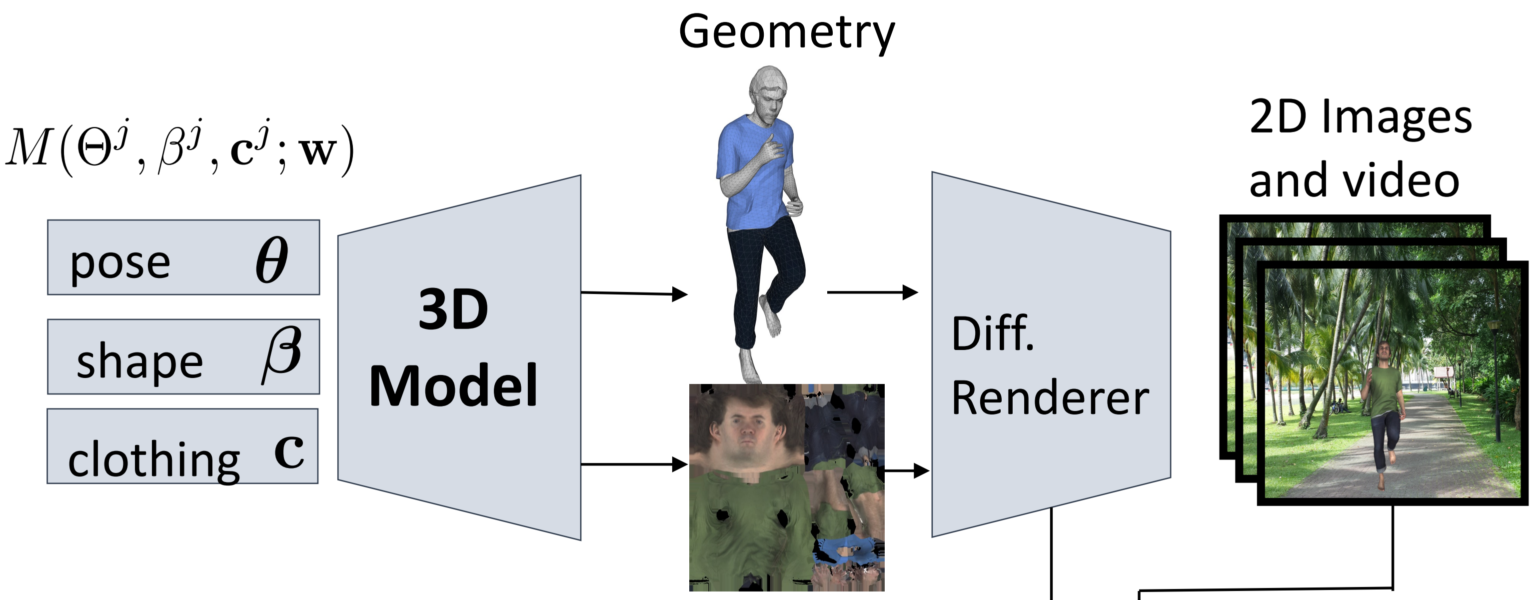
–**No feedback loop**

Connections between **model-based optimization** and **regression** based methods

Capture and learning models in the wild

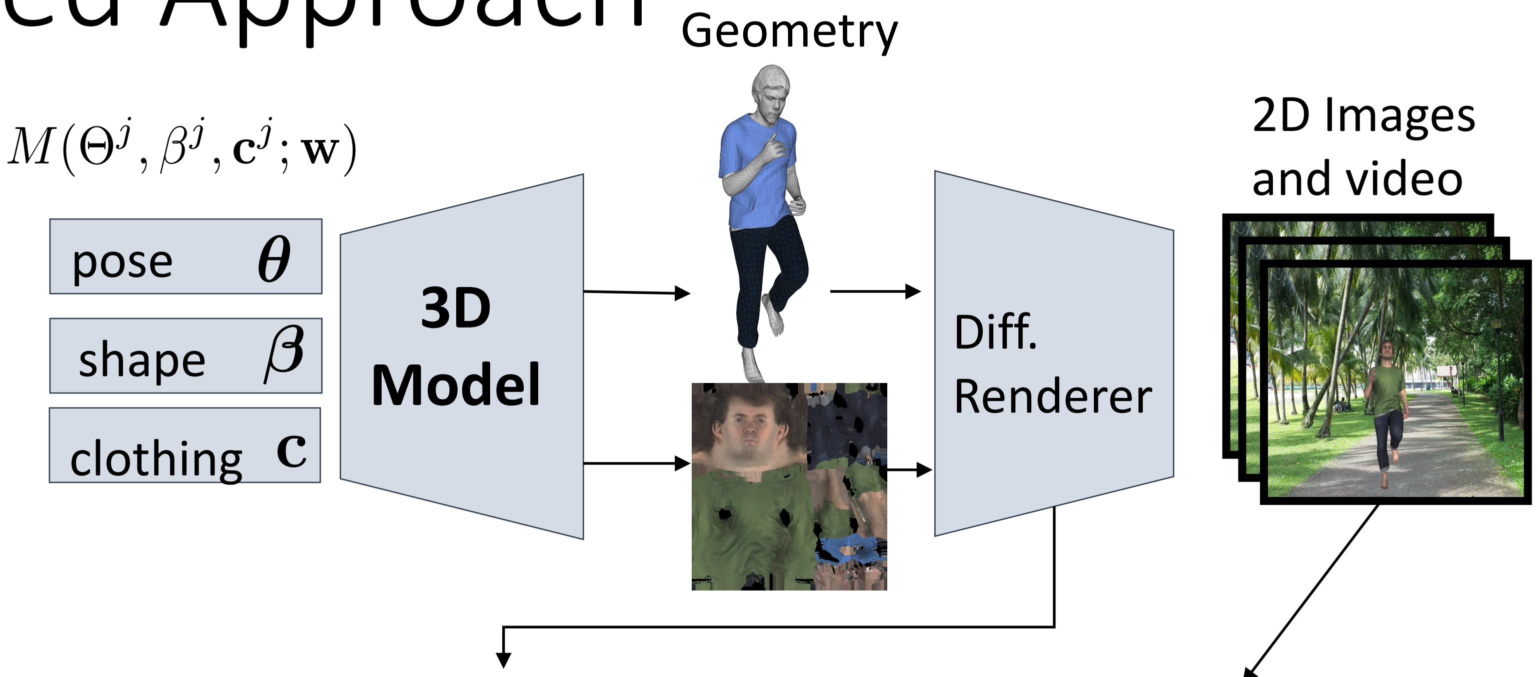


Model-Based Approach



$$\arg\min_{\theta,\beta,\mathbf{c}} \mathrm{dist}(R(M(\theta,\beta,\mathbf{c})),\mathbf{I})$$

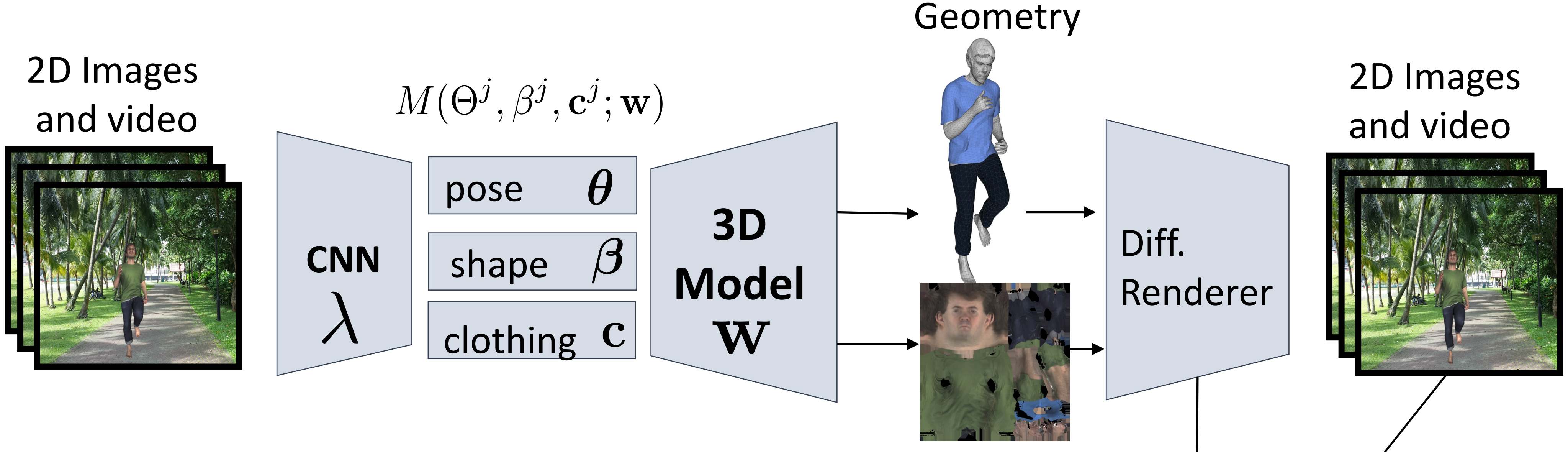
Model-Based Approach



$$\arg\min_{\theta,\beta,\mathbf{c}} \mathrm{dist}(\hat{\mathbf{z}}(R(M(\theta,\beta,\mathbf{c}))), \mathbf{z}(\mathbf{I}))$$

- Slow optimization
- Requires initialization
- Assumes a 3D model is trained

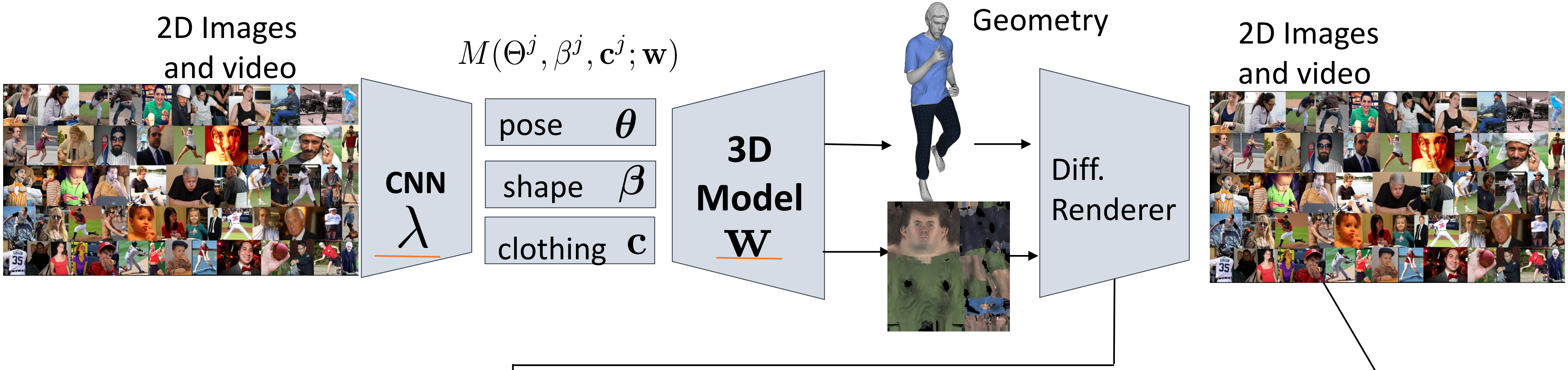
Hybrid Approach (Learning + Model-Based)



$$\arg\min_{\theta,\beta,\mathbf{c}} \mathrm{dist}(\hat{\mathbf{z}}(R(M(\theta,\beta,\mathbf{c}))),\mathbf{z}(\mathbf{I}))$$

$$\theta \mapsto \theta(\mathbf{I};\lambda) \quad \beta \mapsto \beta(\mathbf{I};\lambda) \quad \mathbf{c} \mapsto \mathbf{c}(\mathbf{I};\lambda)$$

Hybrid Approach (Learning + Model-Based)



$$\arg\min_{\lambda,\mathbf{w}} \mathrm{dist}(\hat{\mathbf{z}}(R(M(\theta(\mathbf{I};\lambda),\beta(\mathbf{I};\lambda),\mathbf{c}(\mathbf{I};\lambda);\mathbf{w}))),\mathbf{z}(\mathbf{I}))$$

$$\arg\min_{\lambda,\mathbf{w}} \sum_{I\in\mathcal{D}} \mathrm{dist}(\hat{\mathbf{z}}(R(M(\theta(\mathbf{I}^i;\lambda),\beta(\mathbf{I}^i;\lambda),\mathbf{c}(\mathbf{I}^i;\lambda);\mathbf{w}))),\mathbf{z}(\mathbf{I}^i))$$

Conclusions

- Top down **optimization** based approaches **require initialization and manual tuning** of objective terms.
- Bottom up **learning based** approaches are **automatic** but **not very accurate**.
- Hybrid methods combine optimization and learning to learn in a **self-supervised** manner.
- Given limited data, **abstract the appearance** (e.g., segmentation, keypoints) for robust training.
- **A small amount of 3D annotations are enough** when used in conjunction with 2D annotations