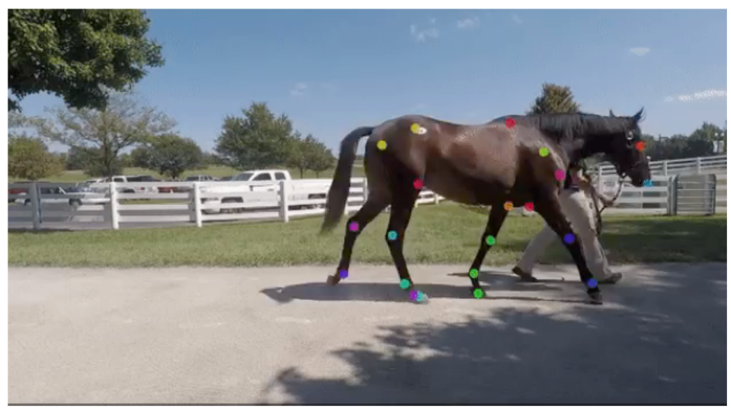
Towards markerless shape and motion capture of animals

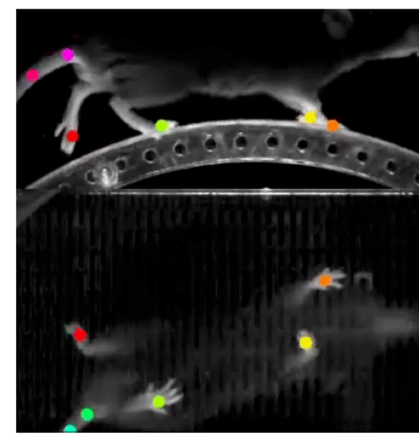
Silvia Zuffi, Angjoo Kanazawa, Michael J. Black

Motion capture of animals

Nature, Oct 2019

- Semi-automatic methods for 2D joints tracking
- Generic, easy to use
- Behavior analysis





A.Mathis et al., DeepLabCut: markerless pose estimation of user-defined body parts with deep learning, Nature Neuroscience, 2018

Motion capture of animals

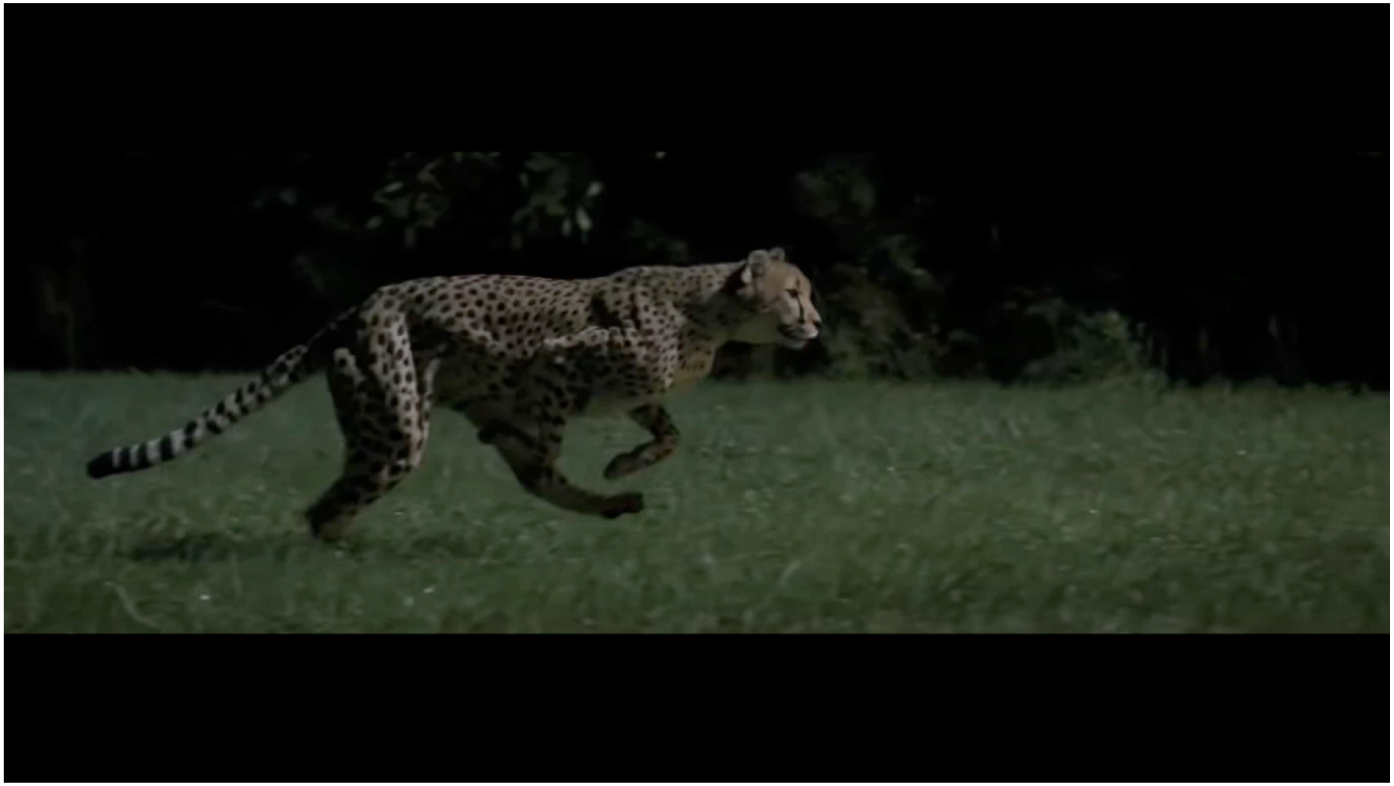
- 3D marker-based systems
- Specific, require trained animals
- Biomechanical studies, animation

Animatrik

CAMERA

Animal markerless mocap

Goal: 3D motion capture of wild animals
+ shape



Animal shape capture

♀ ♂

Animal shape capture

Human markerless mocap

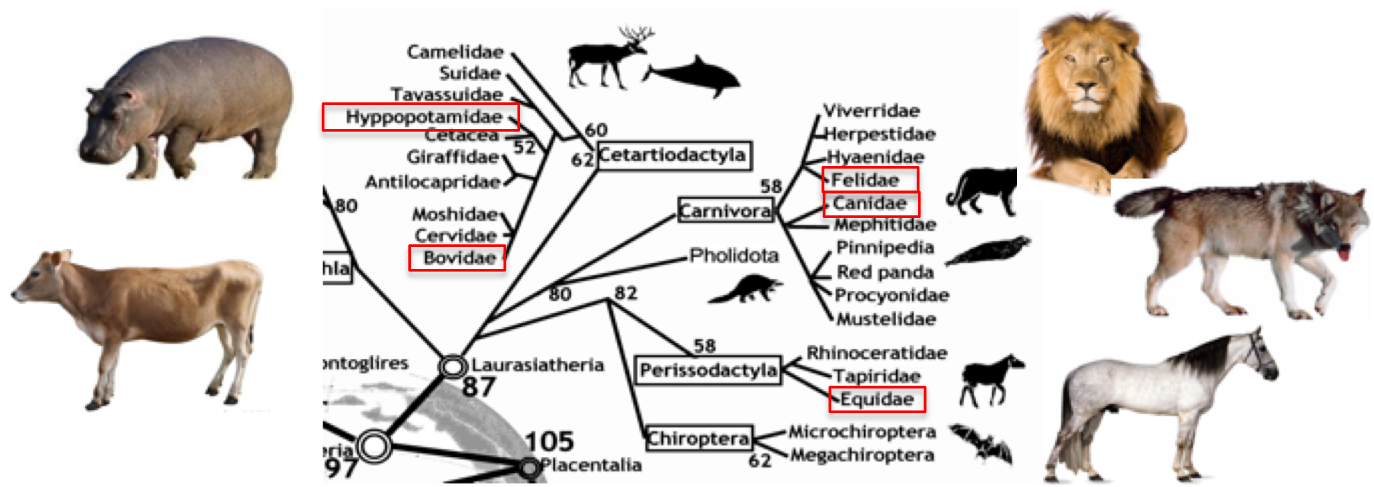
N. Kolotouros, G. Pavlakos, M. J. Black, K. Daniilidis, Learning to Reconstruct 3D Human Pose and Shape via Model-fitting in the Loop, ICCV2019

A. Kanazawa, J. Y. Zhang, P. Felsen, J. Malik, Learning 3D Human Dynamics from Video, CVPR2019

M. Loper et al., SMPL: a Skinned Multi-Person Linear Model, Siggraph2015

SMAL

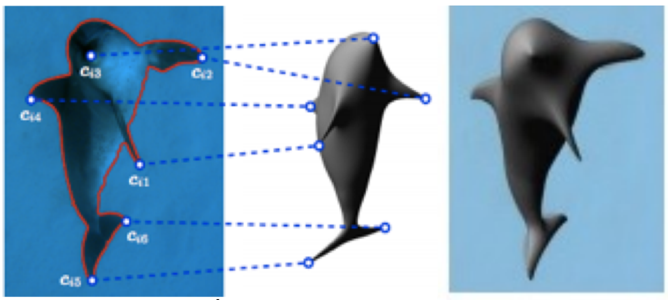
- **S**kinned **M**ulti-**A**nimal **L**inear model
- A 3D shape model representing **articulation** and **shape variation** across different species



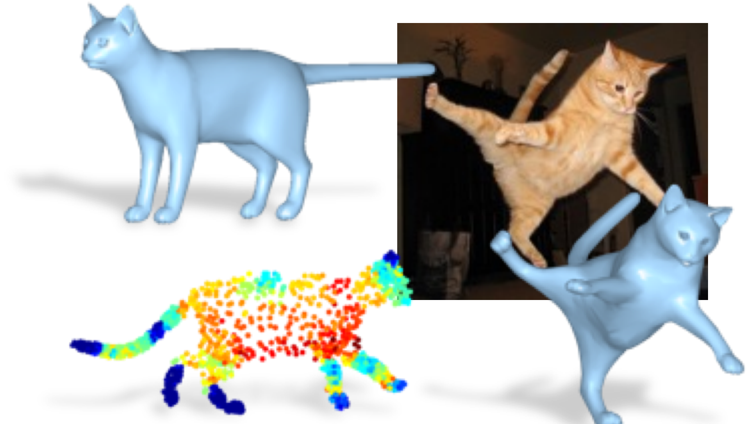
- **From 3D data**

S. Zuffi, A. Kanazawa, D. Jacobs, M.J. Black, 3D Menagerie: Modeling the 3D Shape and Pose of Animals, CVPR 2017

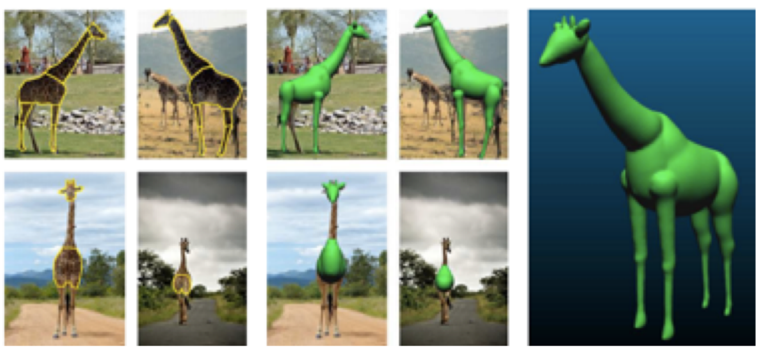
3D animals models: past work



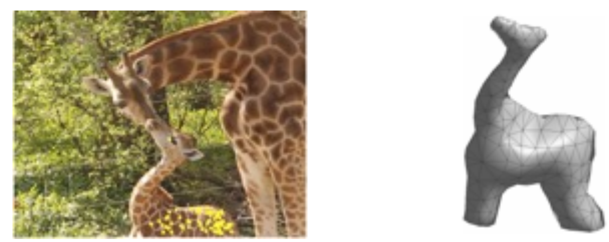
Learn a 3D shape space from images
[Cashman and Fitzgibbon 2012]



Learning 3D deformation of animals
[Kanazawa et al. 2016]



Component-wise modeling
[Ntouskos et al. 2015]



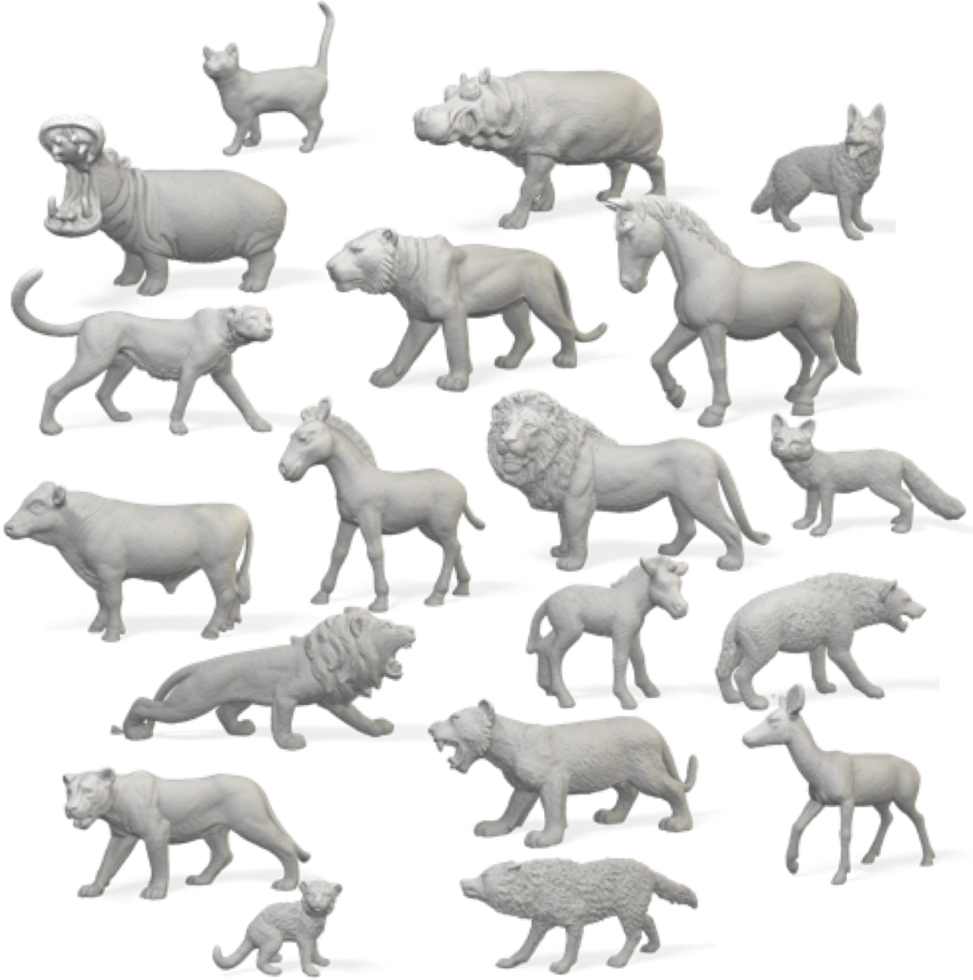
Balloon shapes
[Vincente and Agapito 2013]

None of these learned from 3D and designed with the goal of being a tool for pose and shape estimation from images

Training set

Taxidermy: over smooth, hard to handle ("do not touch!"), not accurate

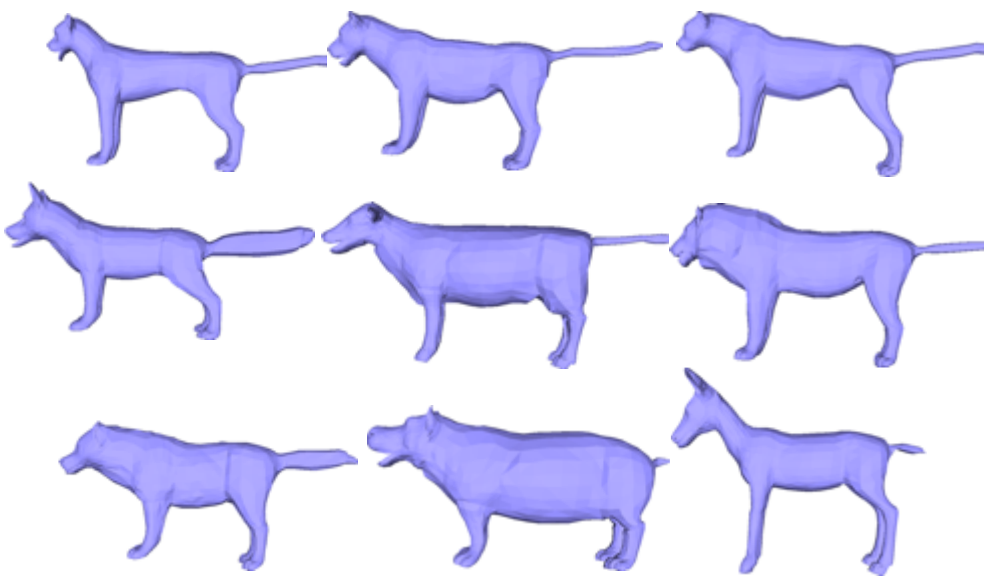
Toys: detailed, easy to get, handle and scan



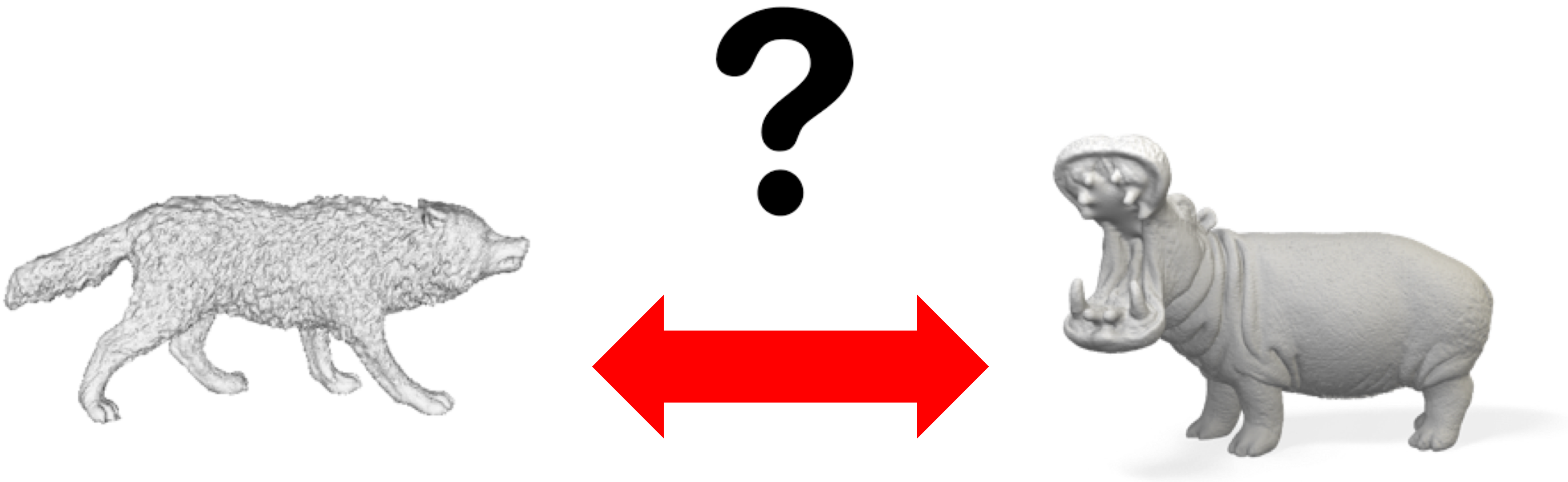
Requirements for learning a 3D articulated shape model

Per-vertex correspondence

All in a reference pose



How to align a wolf to a hippo?

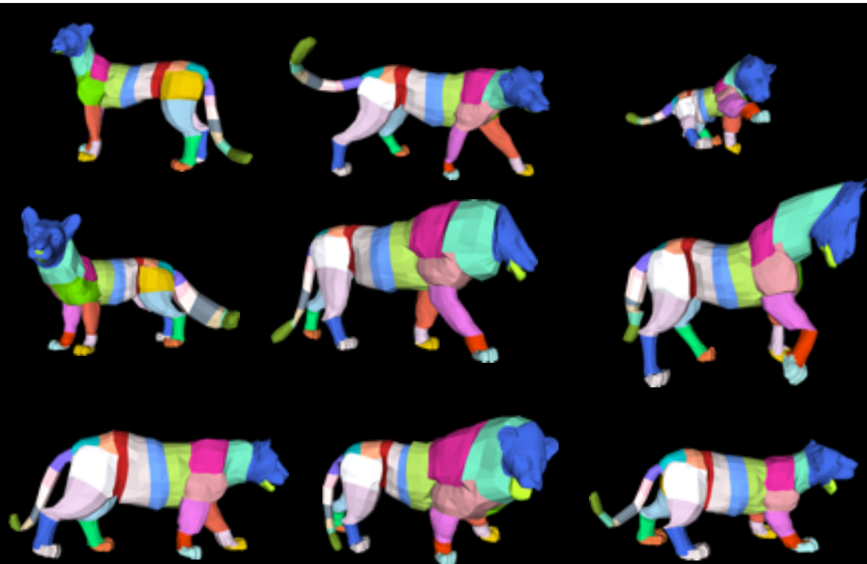


The have different shape and pose!

GLoSS registration

1. Model-based registration: obtain pose estimate and shape approximation

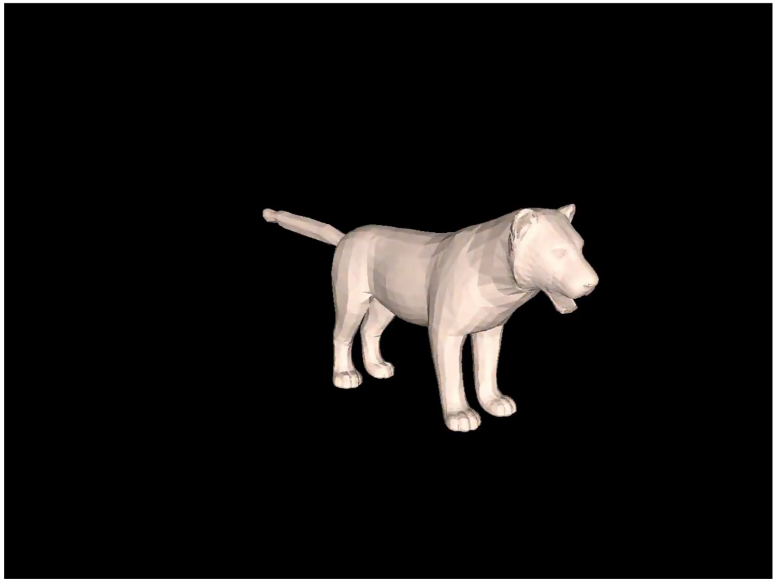
2. Model-free registration: obtain accurate shape and correspondence

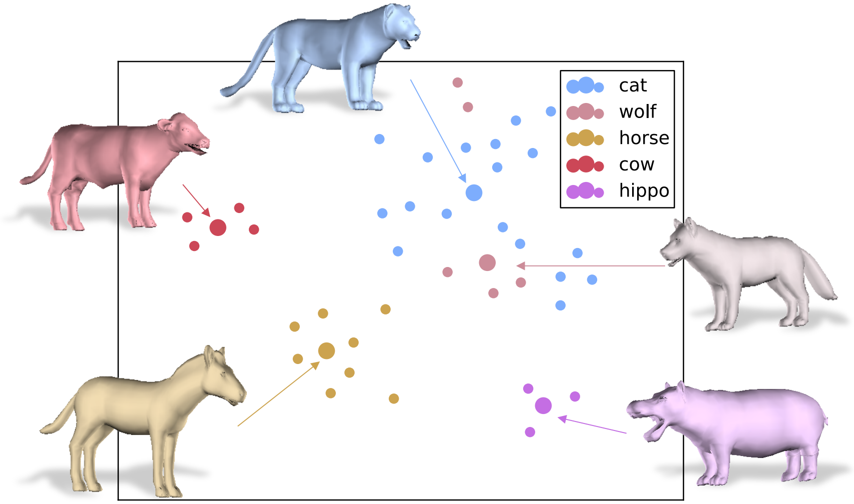


Obtain toys in T-pose

SMAL shape space

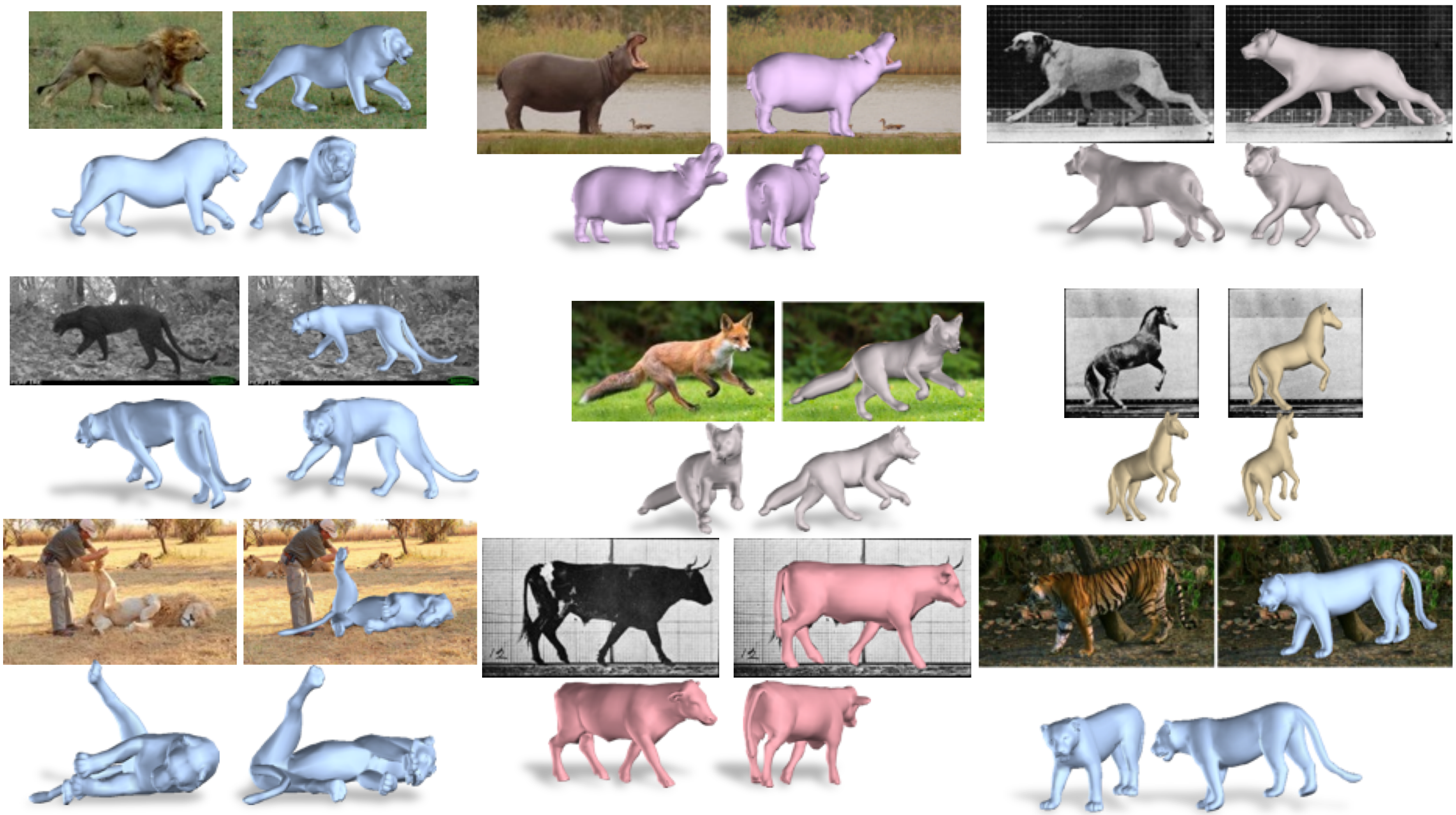
$$\mathbf{v}_{shape}(\beta) = \mathbf{v}_{template} + B_s\beta$$





Fit to images

Manual segmentation and manually annotated keypoints



Fit to video

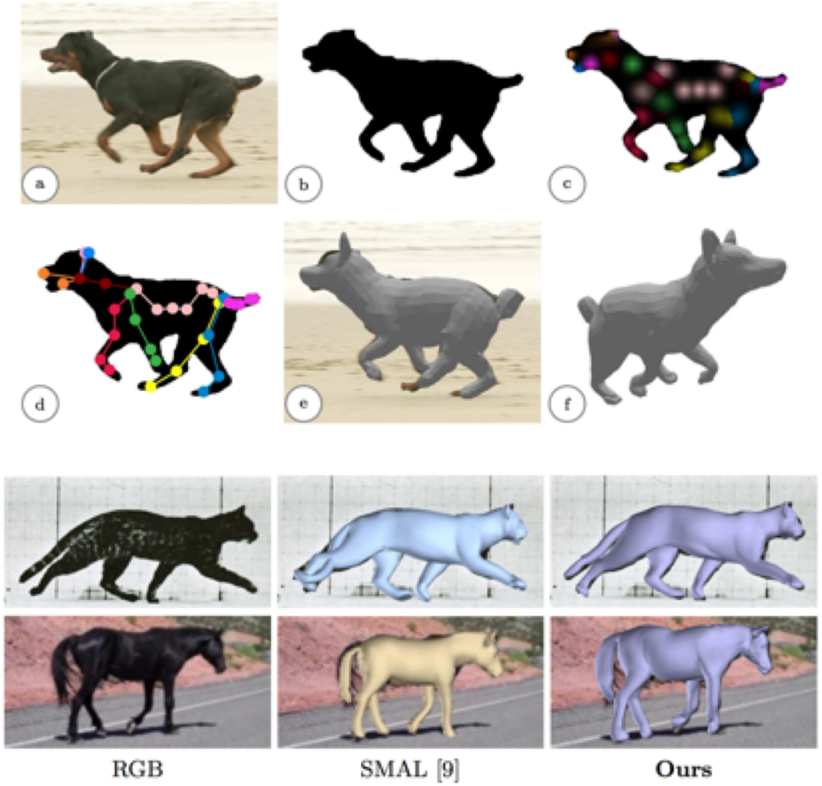
Automatic segmentation and manually annotated keypoints

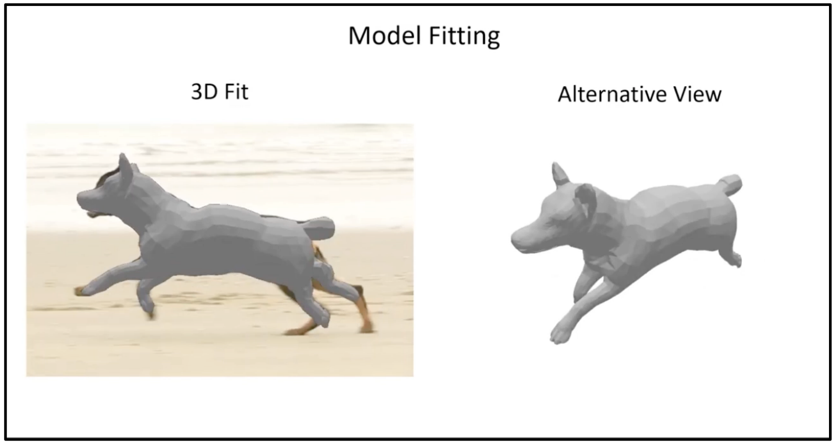
Real cheetah

SMAL fit

Application of SMAL

Automatic segmentation and keypoints detection from silhouette





B. Biggs, T. Roddick, A. Fitzgibbon, R. Cipolla, Creatures great and SMAL: Recovering the shape and motion of animals from video, ACCV2019

Estimate pose and shape from images "in the wild"

- Direct regression from RGB
- Supervised, training based only on synthetic data

S. Zuffi, A. Kanazawa, T. Berger-Wolf, M.J. Black, 3D Safari: Learning to Estimate Zebra Pose, Shape, and Texture from Images "In the Wild", ICCV 2019

Estimate of zebra pose, shape and texture from images

- Predict texture:
  - Hypothesis: predicting texture helps in the task of pose and shape estimation

©Julien Tabet

The Grevy's zebra

RED LIST

ENDANGERED

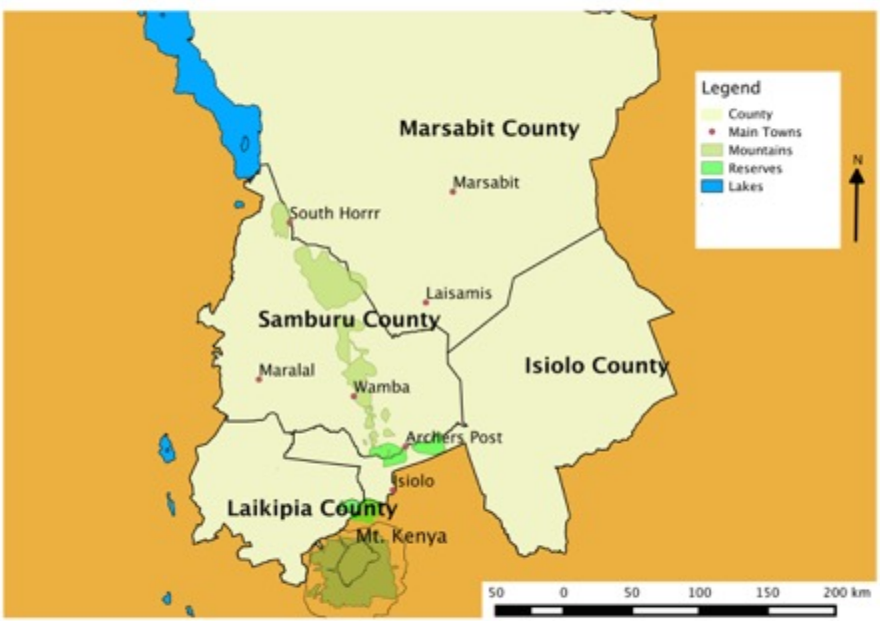
EN

S. Zuffi, A. Kanazawa, T. Berger-Wolf, M.J. Black, 3D Safari: Learning to Estimate Zebra Pose, Shape, and Texture from Images "In the Wild", ICCV 2019

The Grevy's zebra

https://zebra.wildbook.org/

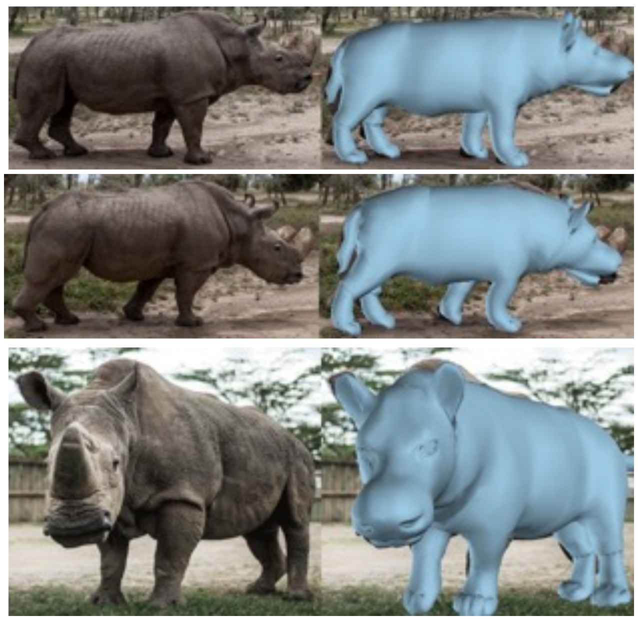
First census of the Grevy's zebra with photographs of ordinary citizens



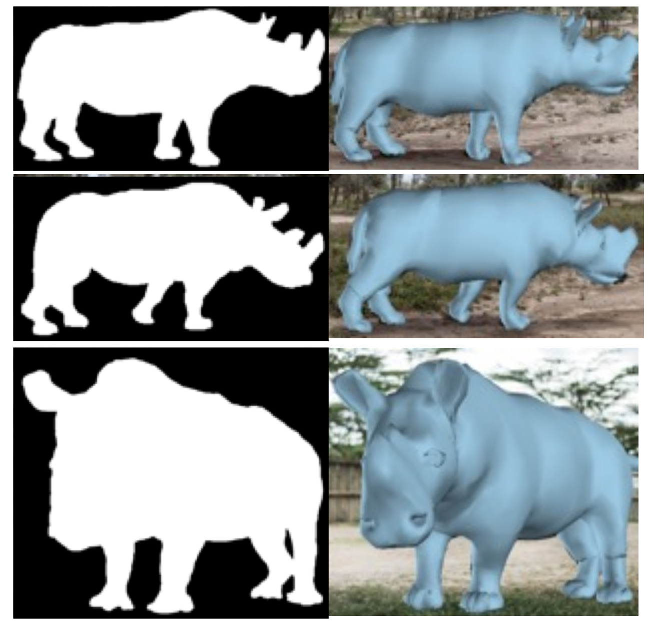
Mpala Research Center, Kenya

SMAL+Refinement (SMALR)

1. SMAL model fitting



2. Model-free shape Refinement

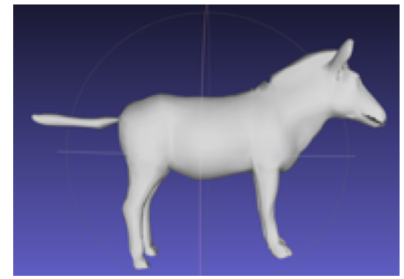


S. Zuffi, A. Kanazawa, M.J.Black, Lions and Tigers and Bears: Capturing Non-Rigid, 3D, Articulated Shape from Images, CVPR2018

Animal avatars with SMALR

Grevy’s zebra avatars

Multiple images of the same subject



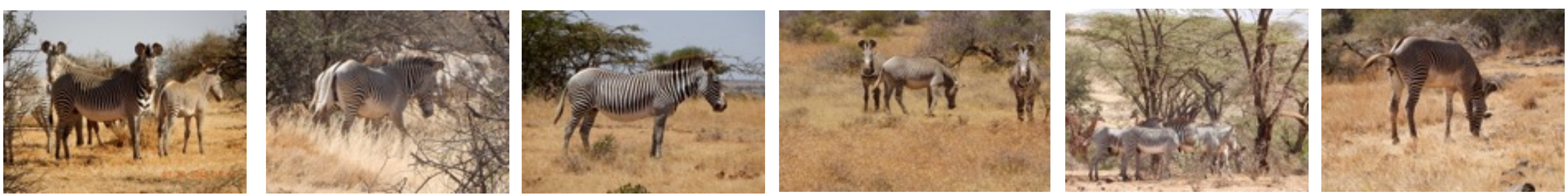
3D model

Texture map

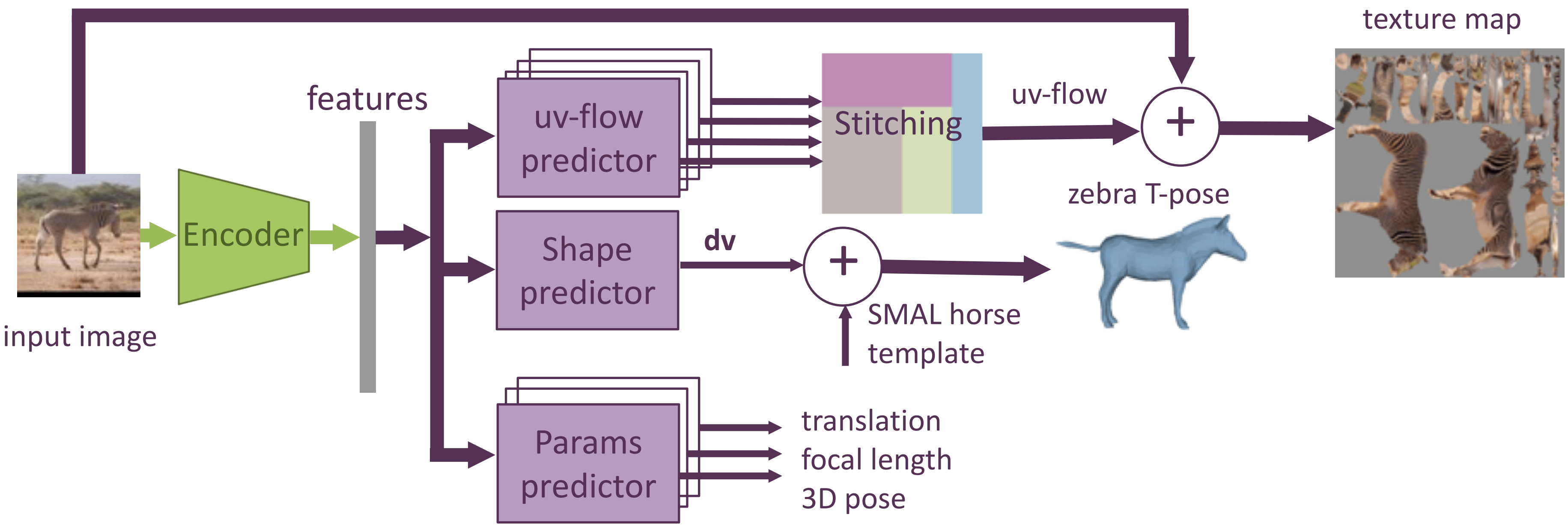
Synthetic dataset from avatars

Synthetic

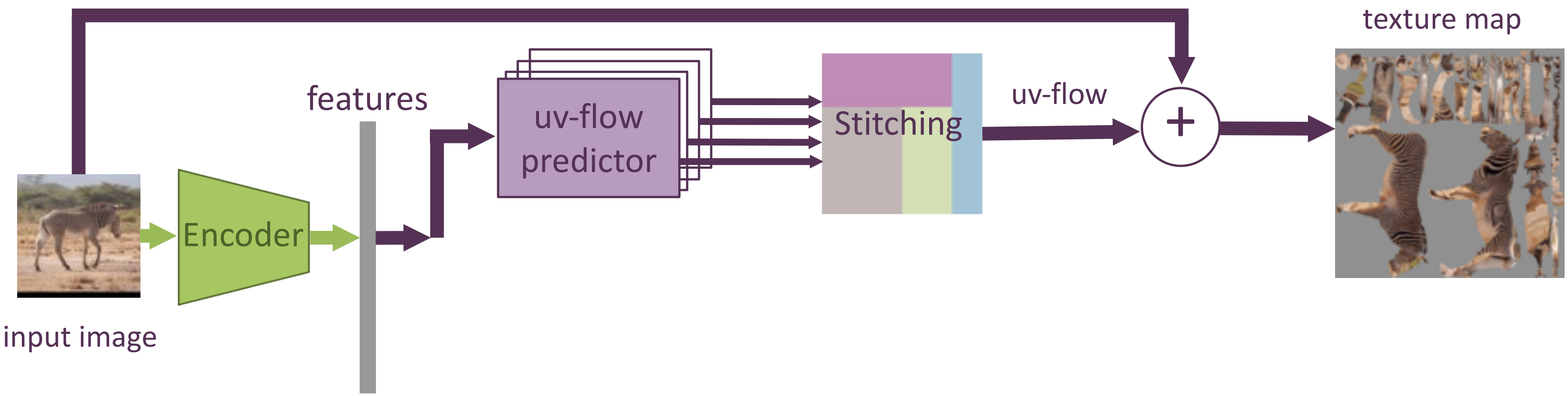
Real

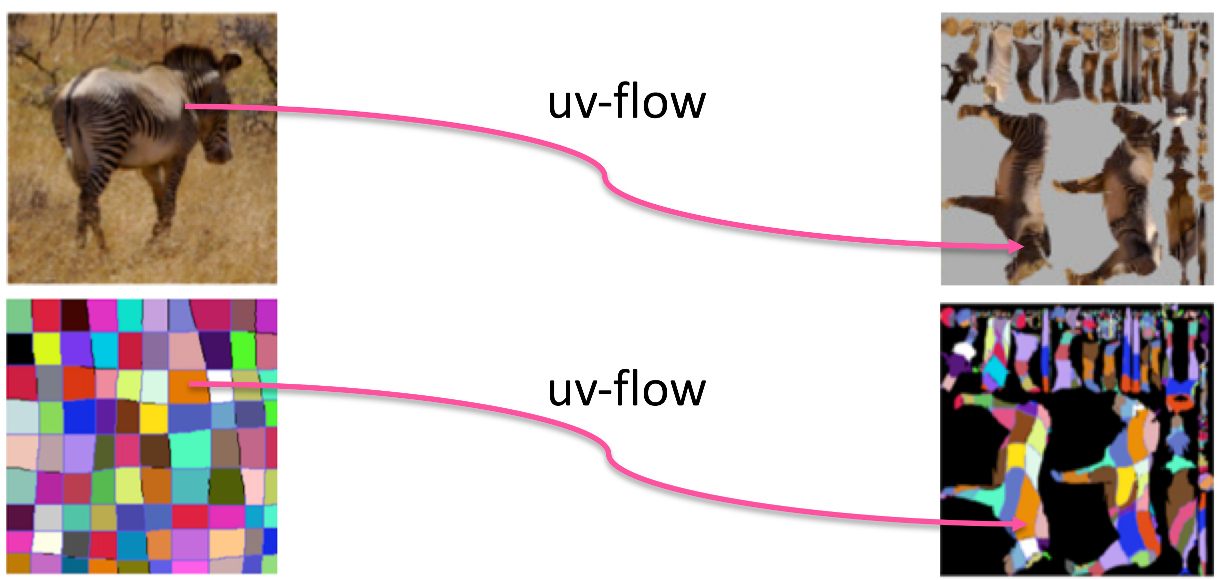


Network

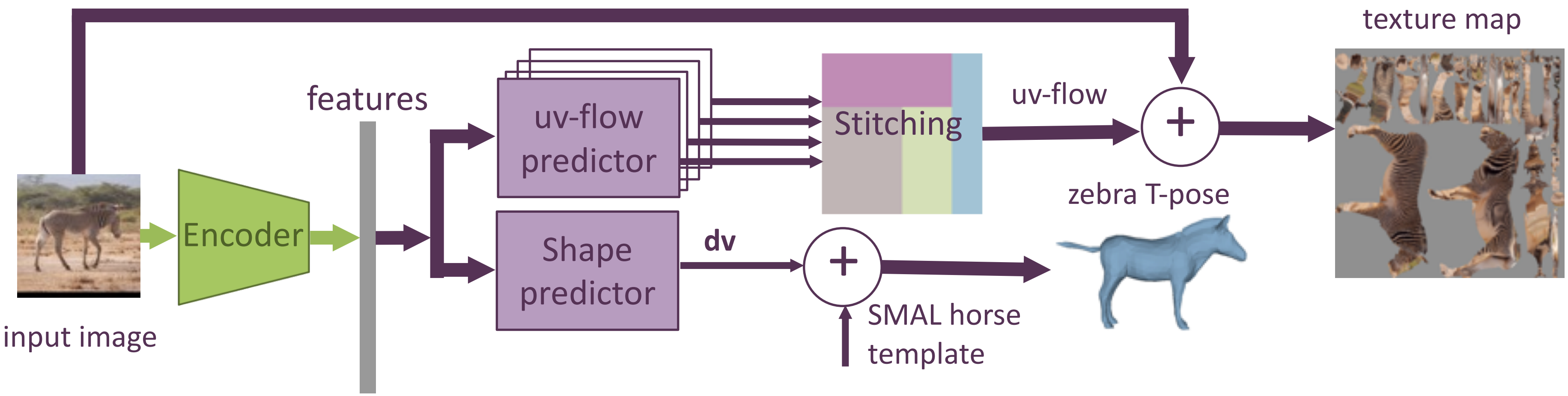

texture map
features
uv-flow
predictor
Stitching
uv-flow
zebra T-pose
Encoder
Shape
predictor
dv
SMAL horse
template
input image
Params
predictor
translation
focal length
3D pose

Network


texture map
features
uv-flow predictor
Stitching
uv-flow
+
Encoder
input image


uv-flow
uv-flow

Network



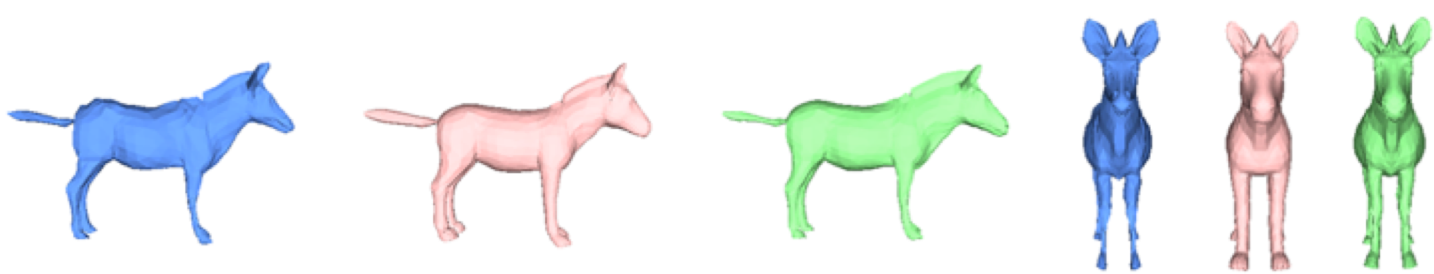
Shape predictor:

$$\mathbf{v}_{shape}(f_s) = \mathbf{v}_{template} + \mathbf{dv}$$

$$\mathbf{dv} = W f_s + b$$

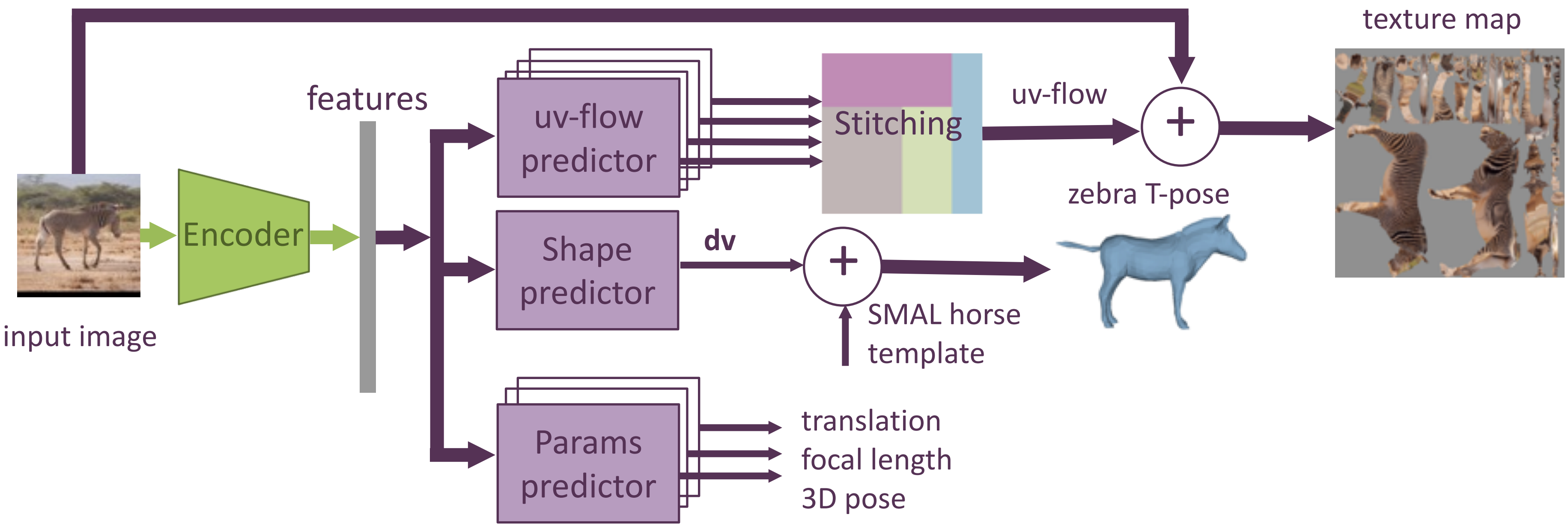
SMAL model:

$$\mathbf{v}_{shape}(\beta) = \mathbf{v}_{template} + B_s \beta$$

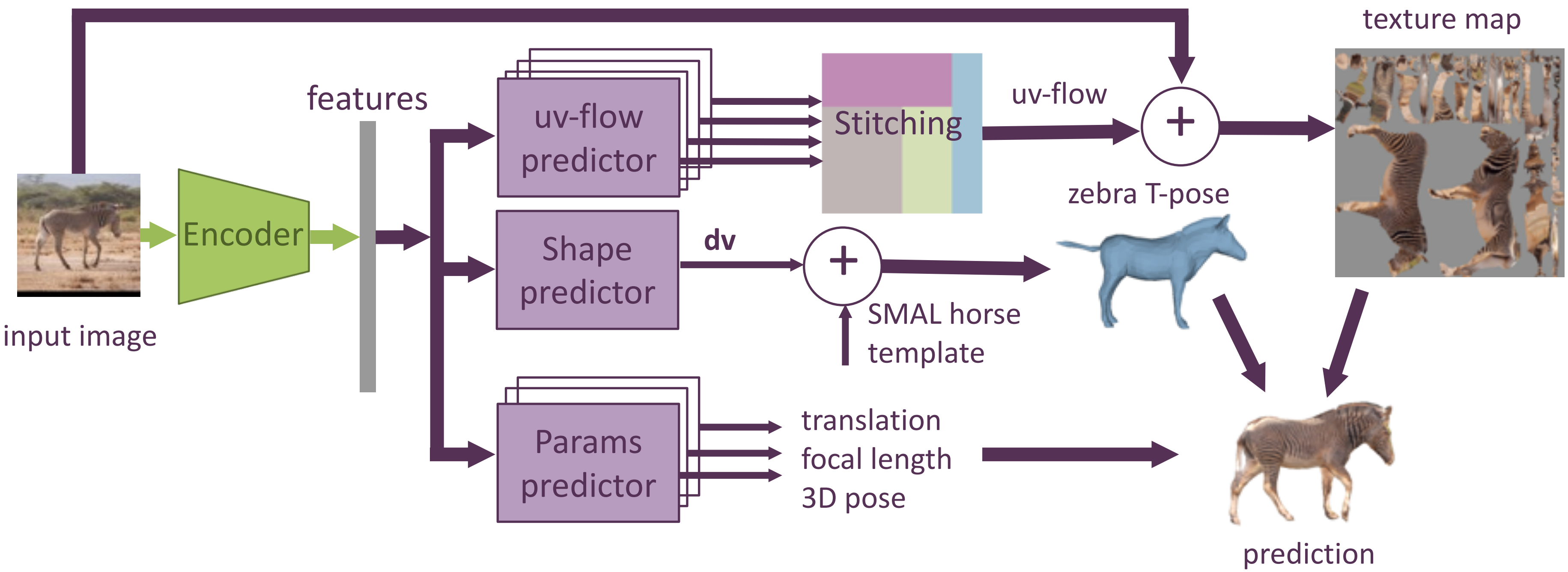


Network “mean“ SMAL Training set mean

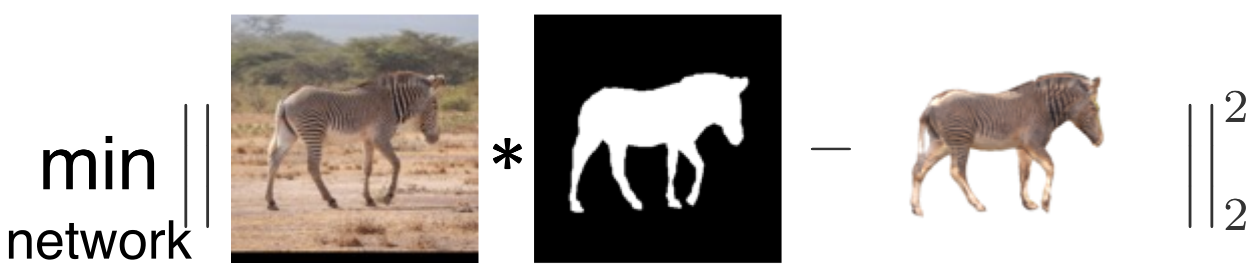
Network


texture map
features
uv-flow predictor
Stitching
uv-flow
zebra T-pose
Encoder
Shape predictor
dv
SMAL horse template
input image
Params predictor
translation
focal length
3D pose

Network

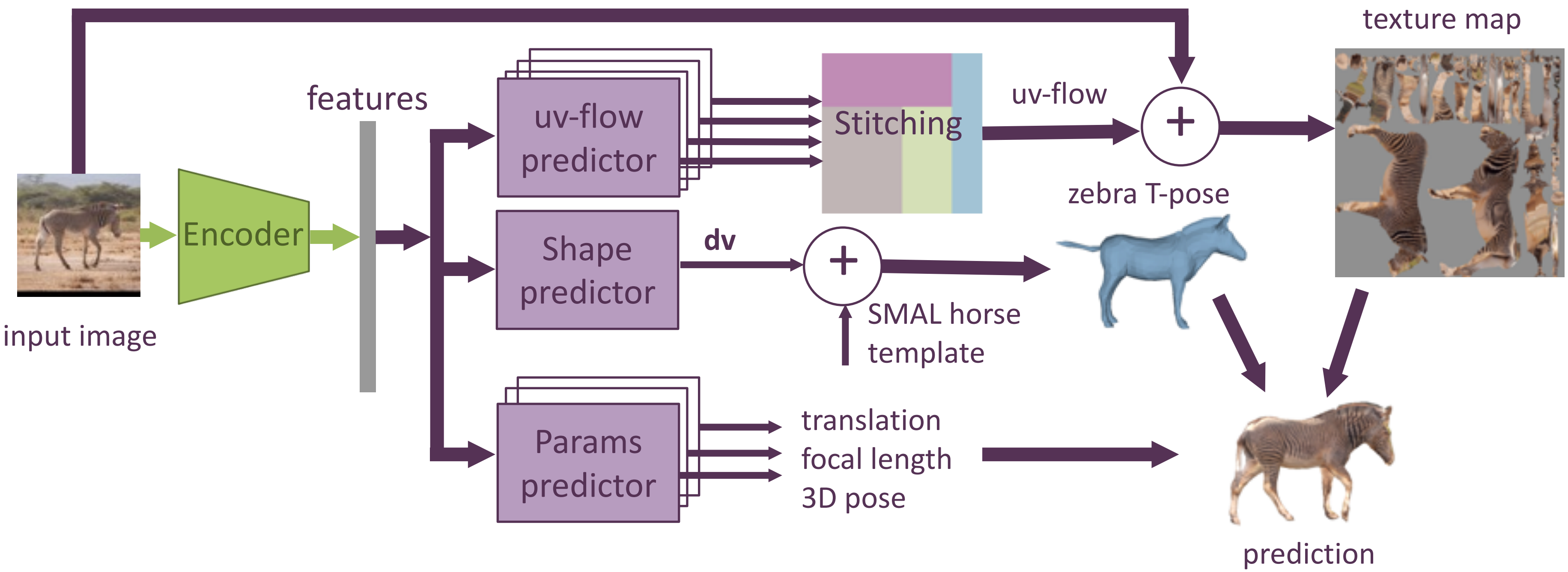


texture map
features
uv-flow predictor
Stitching
uv-flow
zebra T-pose
input image
Encoder
Shape predictor
dv
SMAL horse template
Params predictor
translation
focal length
3D pose
prediction



min
network
*
−
2
2

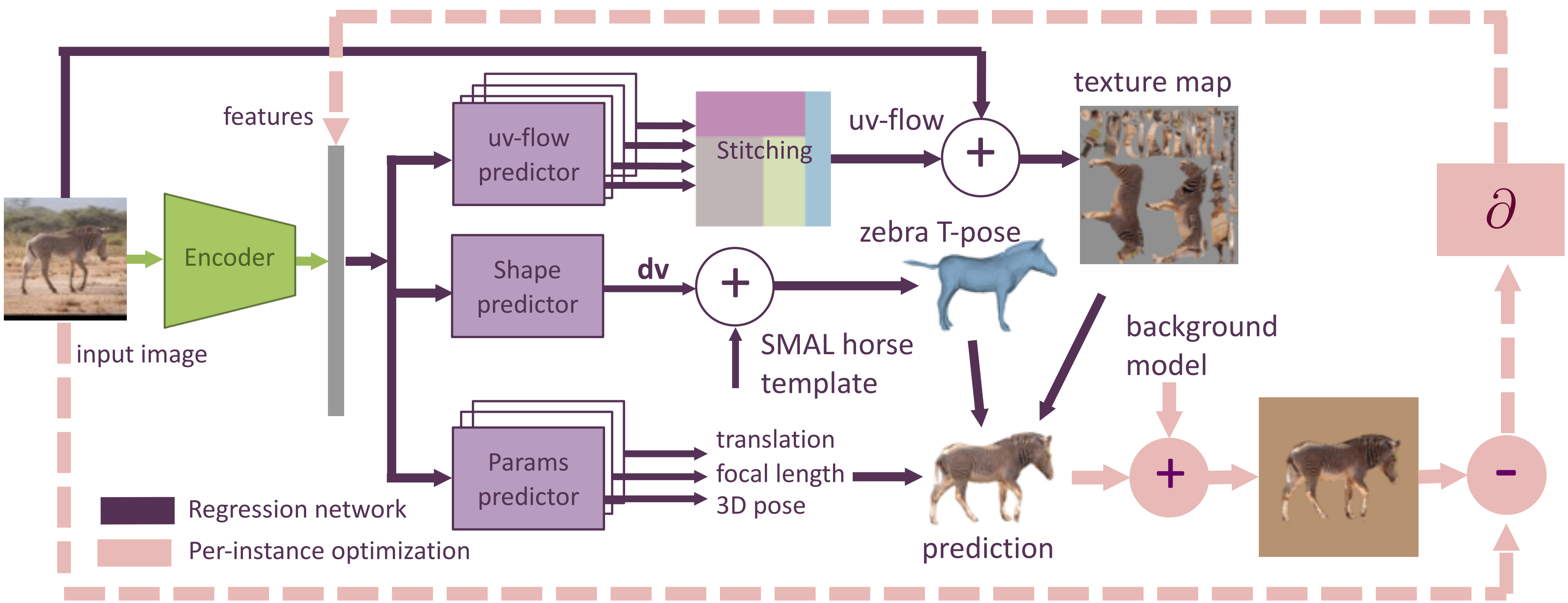
Network

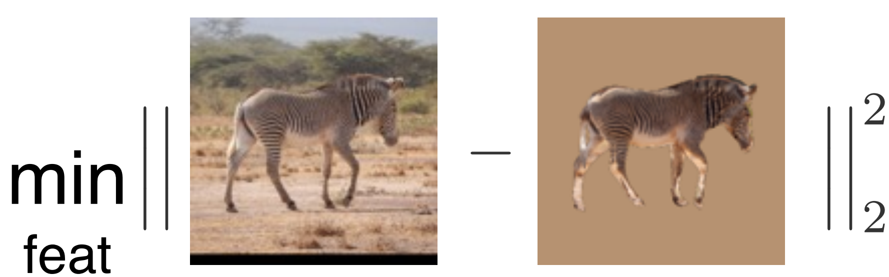


$$L_{train} = L_{mask}(S_{gt}, S) + L_{kp_{2D}}(K_{2D,gt}, K_{2D}) + L_{cam}(f_{gt}, f) + L_{img}(I_{input}, I, S_{gt}) + L_{pose}(\theta_{gt}, \theta) + L_{trans}(\gamma_{gt}, \gamma) + L_{shape}(\mathbf{dv}_{gt}, \mathbf{dv}) + L_{uv}(\mathbf{uv}_{gt}, \mathbf{uv}) + L_{tex}(T_{gt}, T) + L_{dt}(\mathbf{uv}, S_{gt})$$

Results on test set

Unsupervised optimization





Minimize reconstruction loss wrt the latent features, fixing all the decoders

Unsupervised optimization

input	Initial prediction	predicted image	After optimization

Unsupervised optimization resolves small misalignment issues

Results

Method	PCK@0.05	PCK@0.1	IoU
(A) SMAL (gt kp and seg)	92.2	99.4	0.463
(B) feed-forward on synthetic	80.4	97.1	0.423
(C) opt features	**62.3**	**81.6**	**0.422**
(D) opt variables	59.2	80.6	0.418
(E) opt features bg img	59.7	80.5	0.416
(F) feed-forward pred.	59.5	80.3	0.416
(G) no texture	52.3	76.2	0.401
(H) noise bbox	58.7	79.9	0.415

Texture prediction helps!

Better to optimize over features

No texture

With texture

No texture

With texture

Towards markerless shape and motion capture of animals

Silvia Zuffi, Angjoo Kanazawa, Michael J. Black