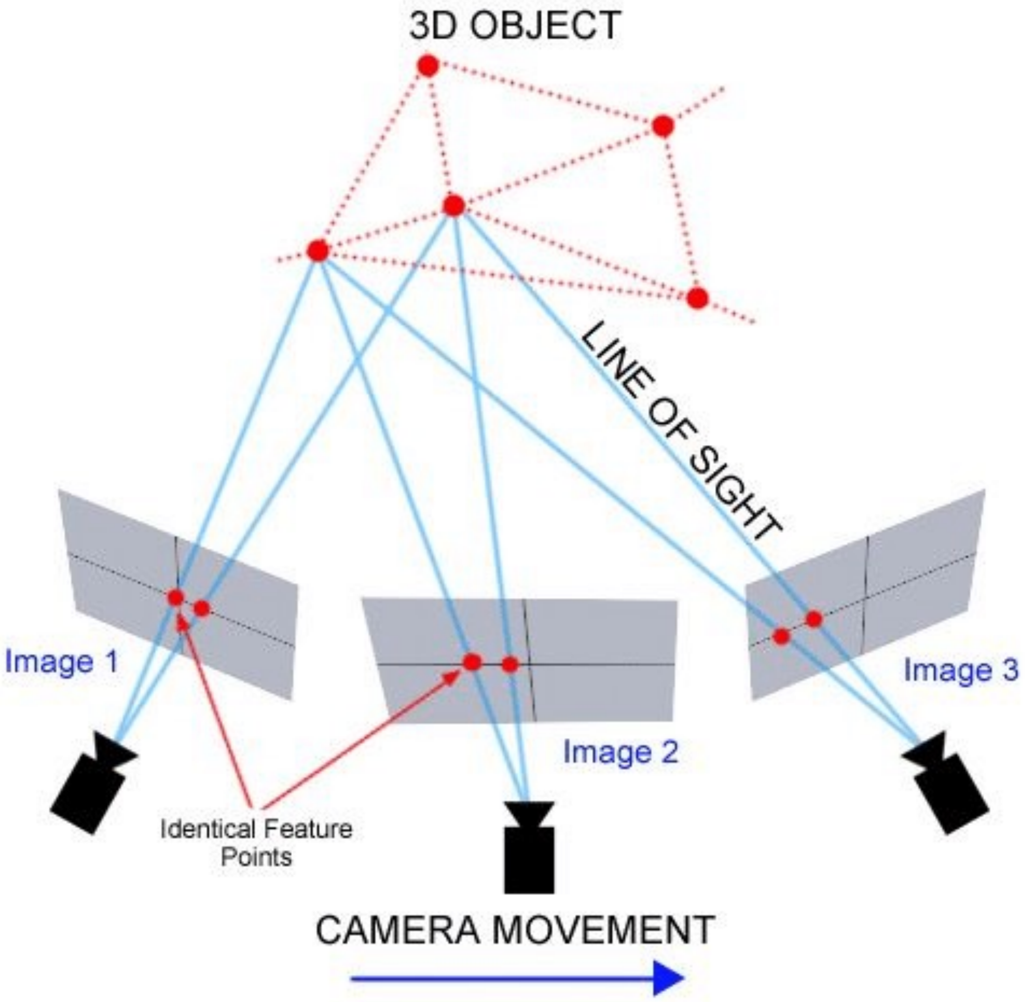
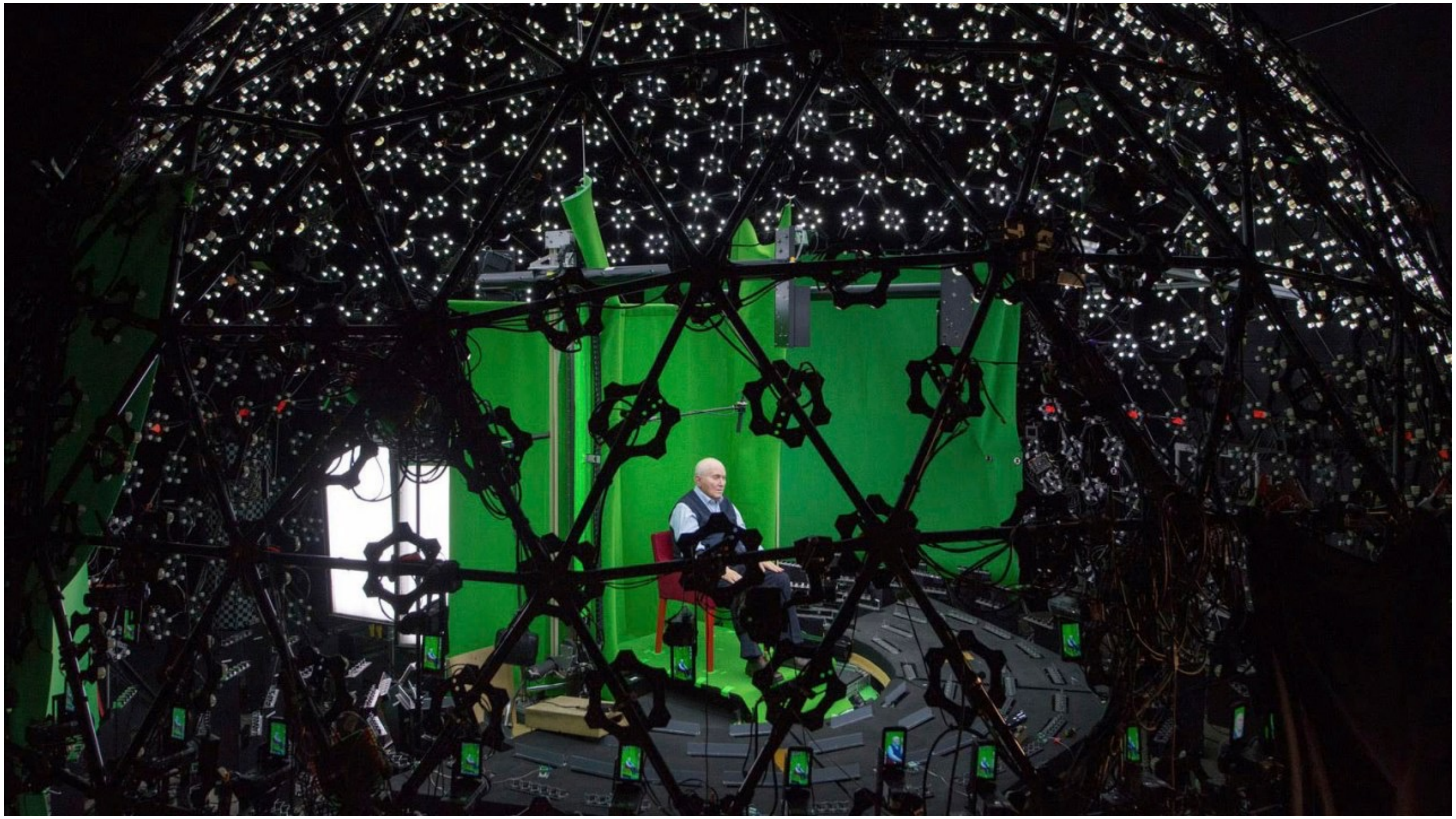
Learning Articulated 3D Animals from Internet Images

UNIVERSITY OF OXFORD

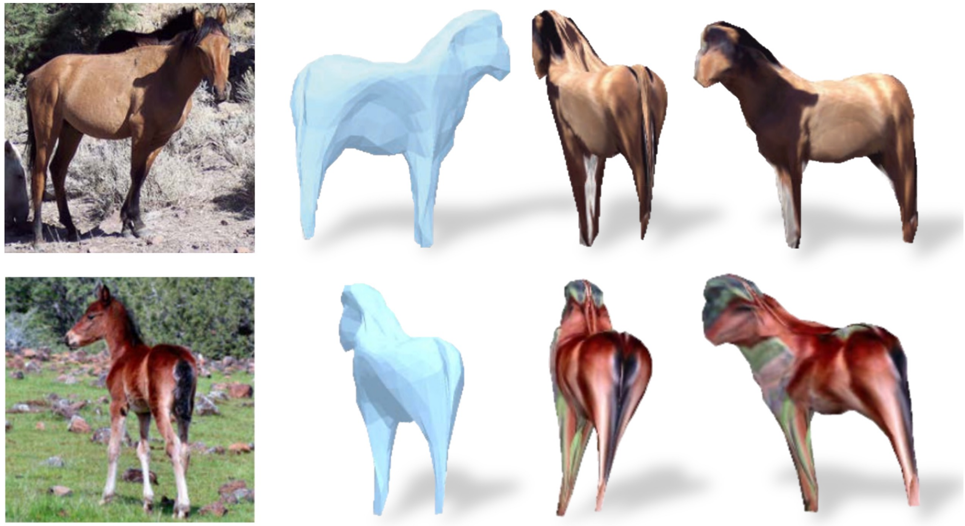
Tomas Jakab, University of Oxford, VGG


3D OBJECT
LINE OF SIGHT
Image 1
Image 2
Image 3
Identical Feature Points
CAMERA MOVEMENT



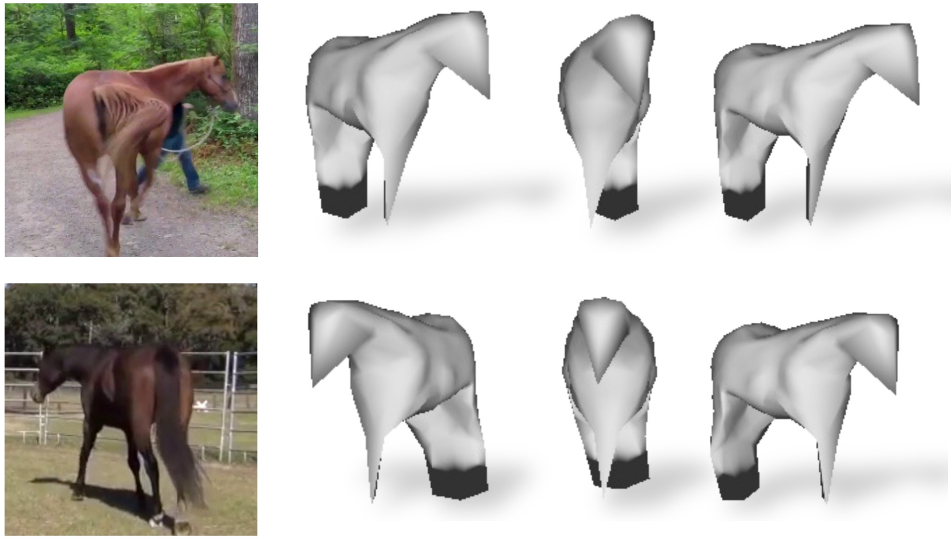
Training Data – Single View Images

Prior work



Input Input view Other views

UMR [1]



Input Input view Other views

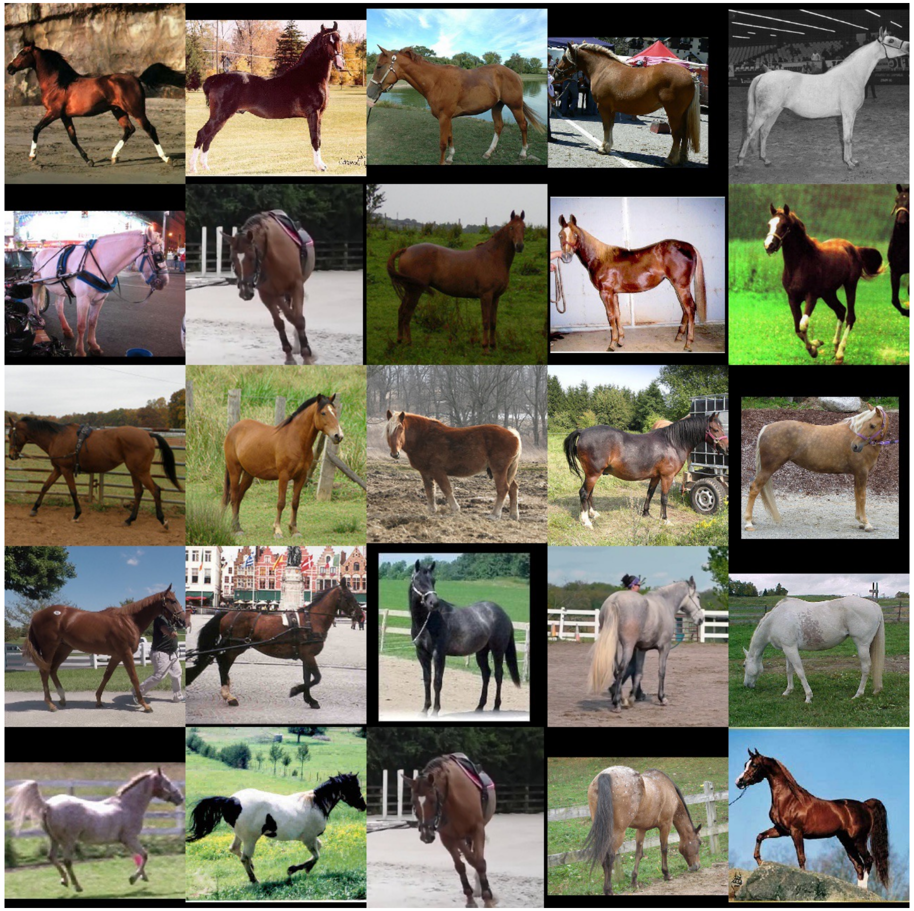
DOVE [2]

[1] Self-supervised single-view 3d reconstruction via semantic consistency.
Li et. al. ECCV 2020.

[2] DOVE: Learning deformable 3d objects by watching videos.
Wu et. al. IJCV, 2023.

Training Data – Single View Images

Training Data – Single View Images

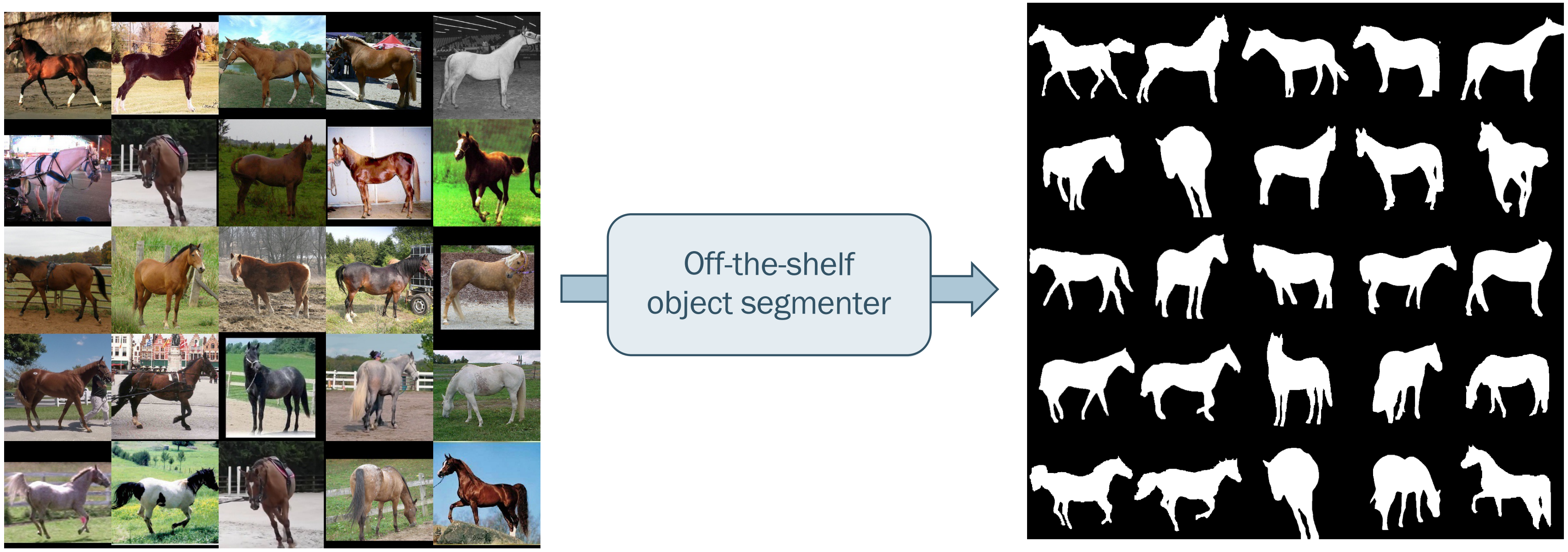


What is different?

Background

Shape

Appearance

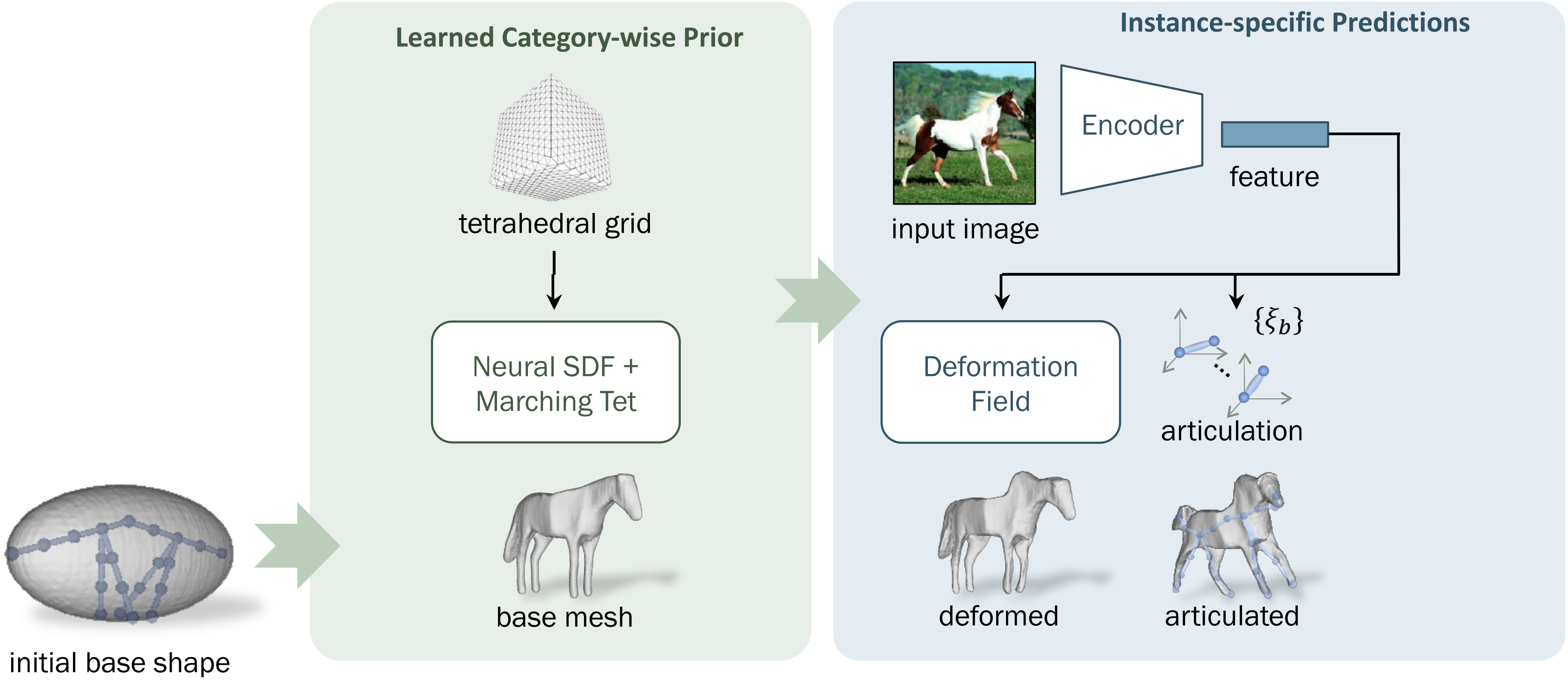


Off-the-shelf object segmenter

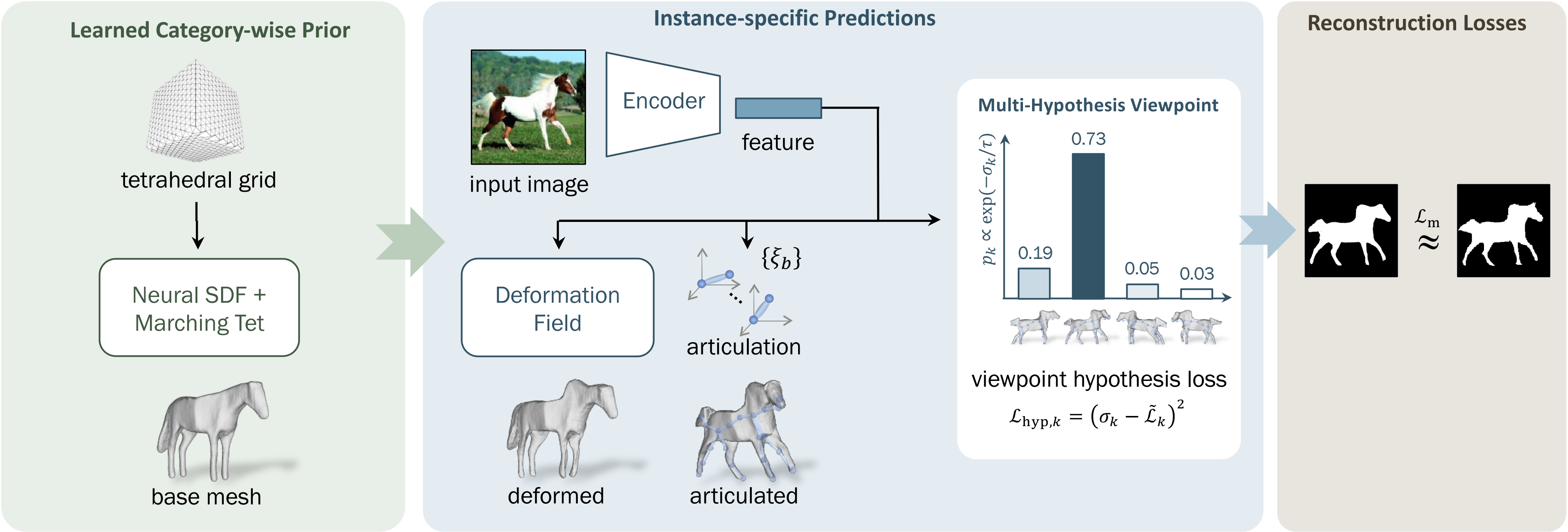
Instance Masks

Modeling shape

Hierarchical Shape Prediction



Camera Prediction



Learned Category-wise Prior

tetrahedral grid

Neural SDF + Marching Tet

base mesh

Instance-specific Predictions

input image

Encoder

feature

Deformation Field

$\{\xi_b\}$

articulation

deformed

articulated

Multi-Hypothesis Viewpoint

$p_k \propto \exp(-\sigma_k/\tau)$

0.19

0.73

0.05

0.03

viewpoint hypothesis loss

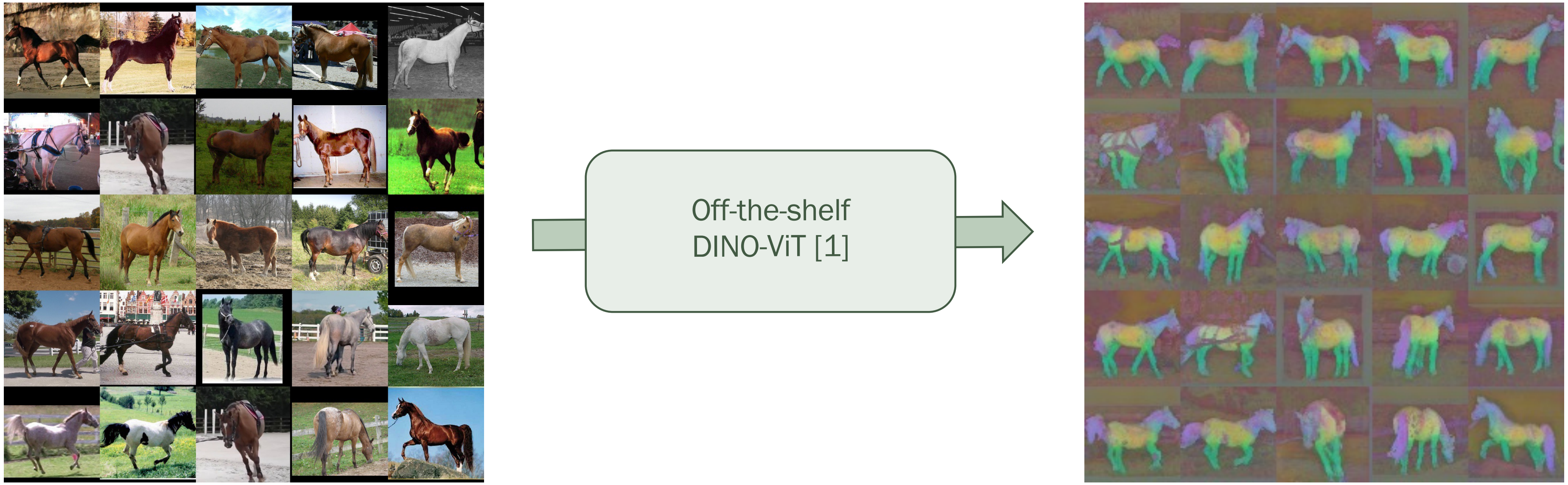
$$\mathcal{L}_{\mathrm{hyp},k} = (\sigma_k - \tilde{\mathcal{L}}_k)^2$$

Reconstruction Losses

$\mathcal{L}_{\mathrm{m}}$

$\approx$

Category Appearance

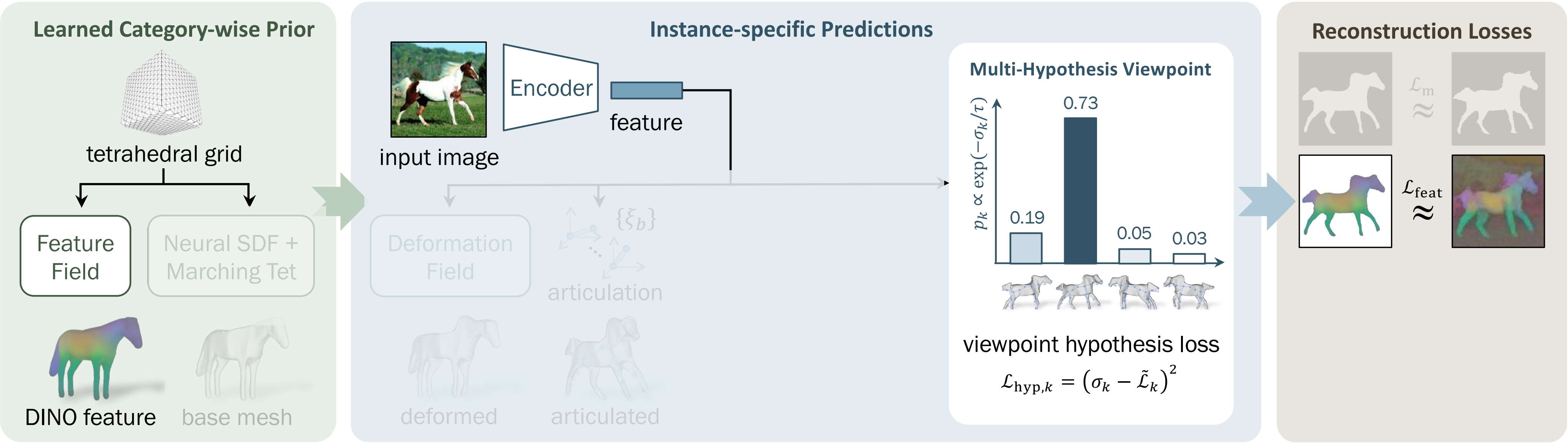


Off-the-shelf
DINO-ViT [1]

Self-supervised Image Features

[1] Emerging Properties in Self-supervised Vision Transformers. Caron et. al. ICCV 2021.

Category Appearance



Learned Category-wise Prior

tetrahedral grid

Feature Field

Neural SDF + Marching Tet

DINO feature

base mesh

Instance-specific Predictions

input image

Encoder

feature

Deformation Field

$\{\xi_b\}$

articulation

deformed

articulated

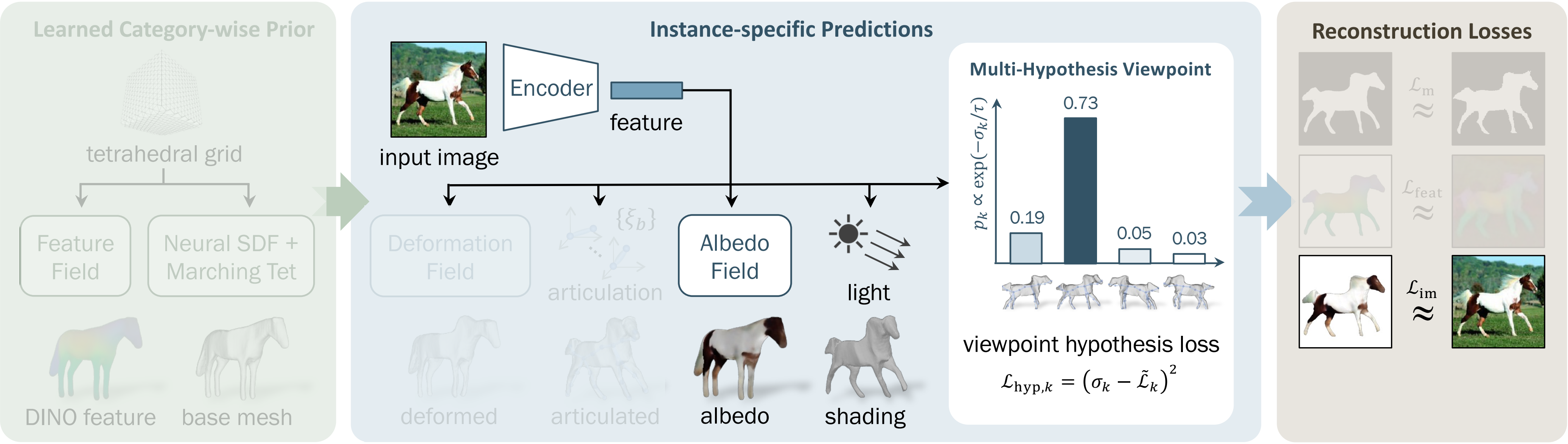
Multi-Hypothesis Viewpoint

viewpoint hypothesis loss

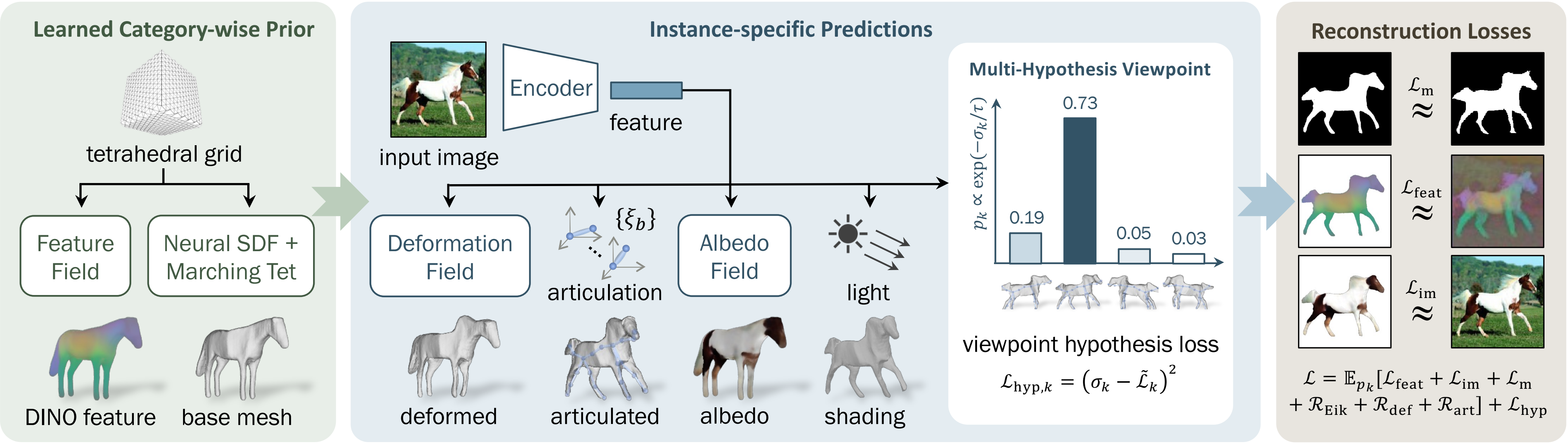
$$\mathcal{L}_{\text{hyp},k} = \left(\sigma_k - \tilde{\mathcal{L}}_k\right)^2$$

Reconstruction Losses

Instance Appearance


Learned Category-wise Prior
tetrahedral grid
Feature Field
Neural SDF + Marching Tet
DINO feature
base mesh
Instance-specific Predictions
Encoder
feature
input image
Deformation Field
$\{\xi_b\}$
articulation
Albedo Field
light
deformed
articulated
albedo
shading
Multi-Hypothesis Viewpoint
$p_k \propto \exp(-\sigma_k/\tau)$
0.73
0.19
0.05
0.03
viewpoint hypothesis loss
$\mathcal{L}_{\text{hyp},k} = (\sigma_k - \tilde{\mathcal{L}}_k)^2$
Reconstruction Losses
$\mathcal{L}_{\text{m}}$
≈
$\mathcal{L}_{\text{feat}}$
≈
$\mathcal{L}_{\text{im}}$
≈

Canonical Appearance



Learned Category-wise Prior

tetrahedral grid

Feature Field

Neural SDF + Marching Tet

DINO feature

base mesh

Instance-specific Predictions

input image

Encoder

feature

Deformation Field

$\{\xi_b\}$

articulation

Albedo Field

light

deformed

articulated

albedo

shading

Multi-Hypothesis Viewpoint

$p_k \propto \exp(-\sigma_k/\tau)$

viewpoint hypothesis loss

$$\mathcal{L}_{\text{hyp},k} = (\sigma_k - \tilde{\mathcal{L}}_k)^2$$

Reconstruction Losses

$\mathcal{L}_{\text{m}}$ $\approx$

$\mathcal{L}_{\text{feat}}$ $\approx$

$\mathcal{L}_{\text{im}}$ $\approx$

$$\mathcal{L} = \mathbb{E}_{p_k}[\mathcal{L}_{\text{feat}} + \mathcal{L}_{\text{im}} + \mathcal{L}_{\text{m}} + \mathcal{R}_{\text{Eik}} + \mathcal{R}_{\text{def}} + \mathcal{R}_{\text{art}}] + \mathcal{L}_{\text{hyp}}$$

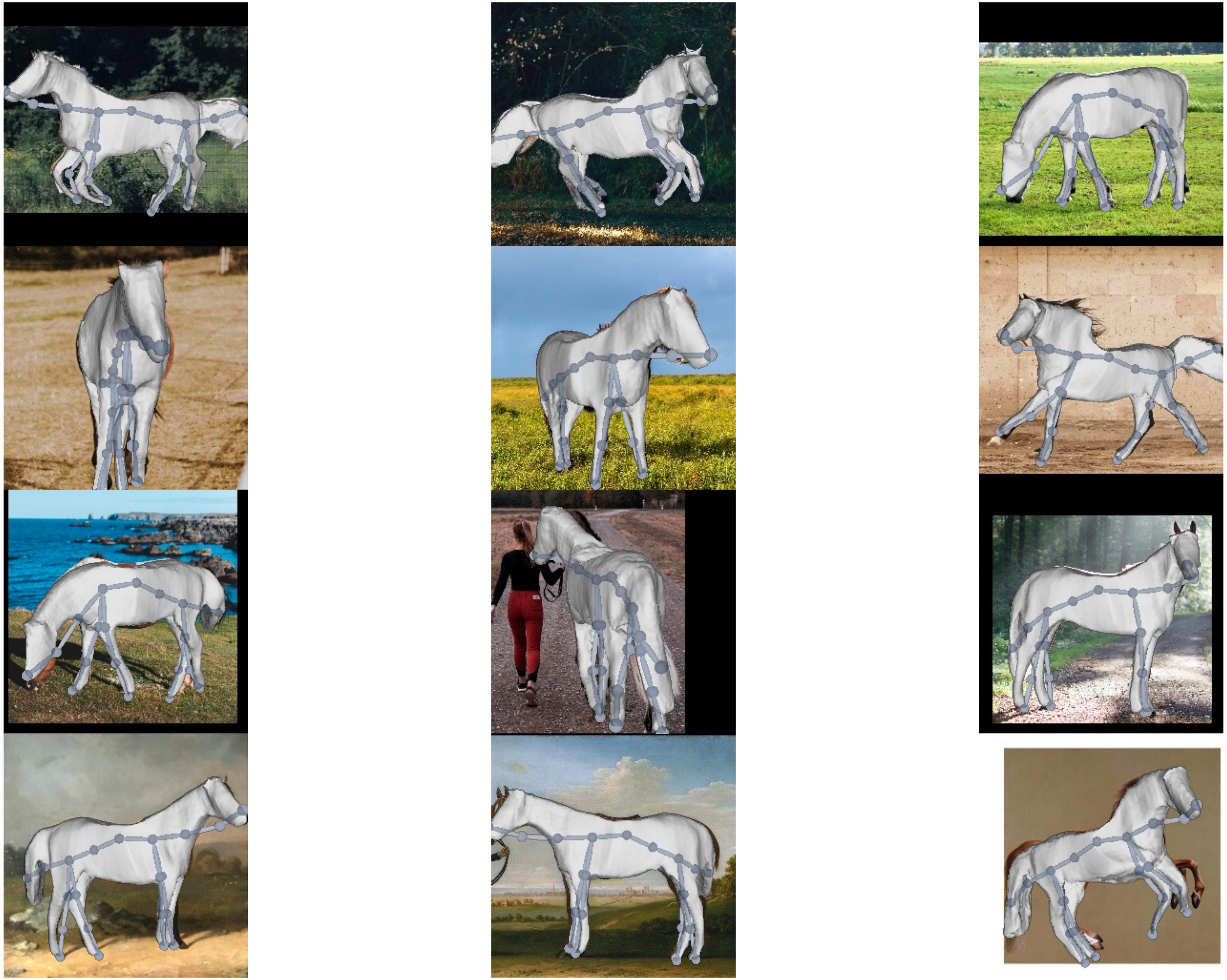
Entire pipeline trained end-to-end with reconstruction losses

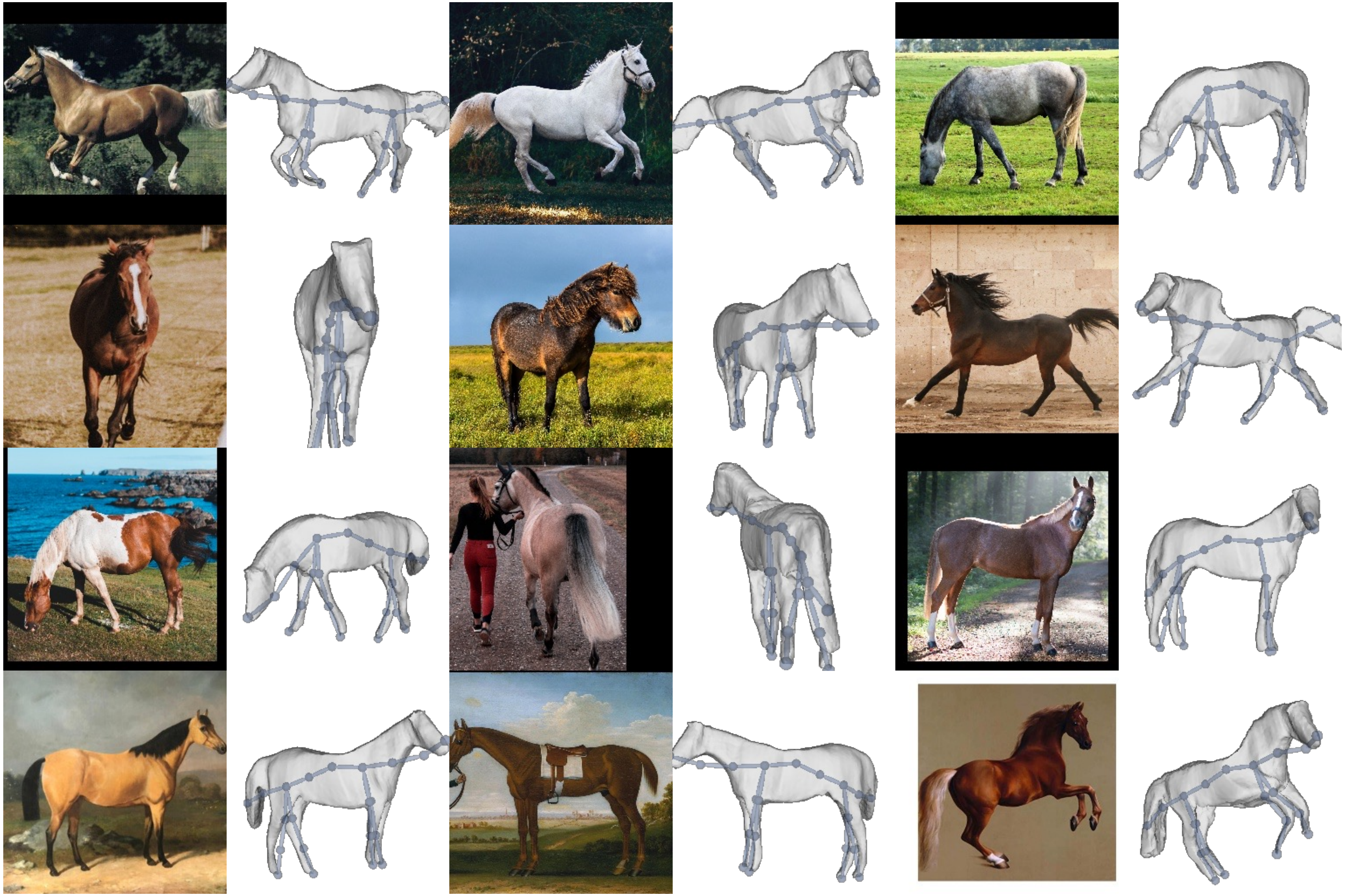
(except for frozen DINO-ViT image encoder, pre-trained via self-supervision)

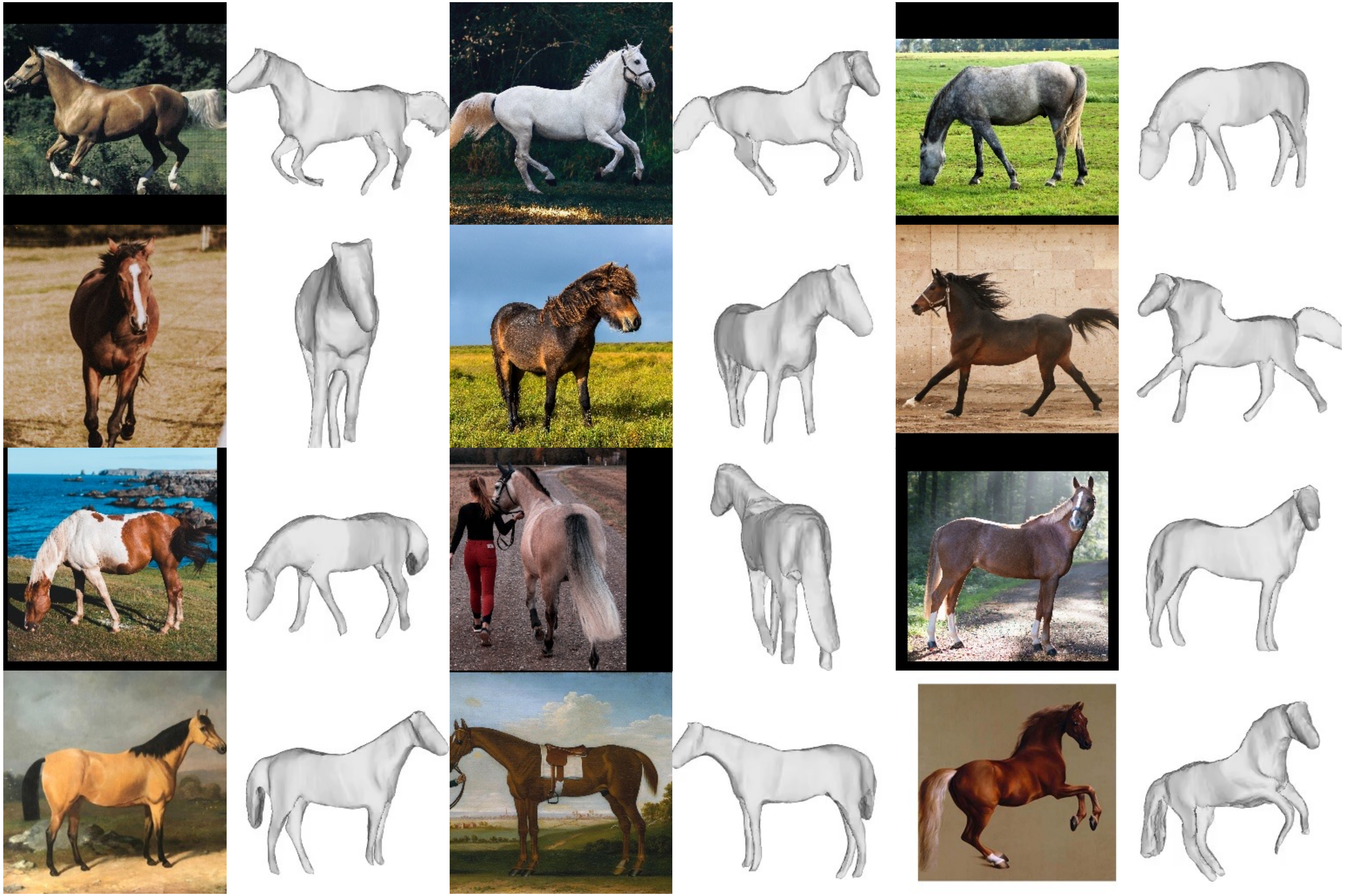
no keypoints, no template shapes

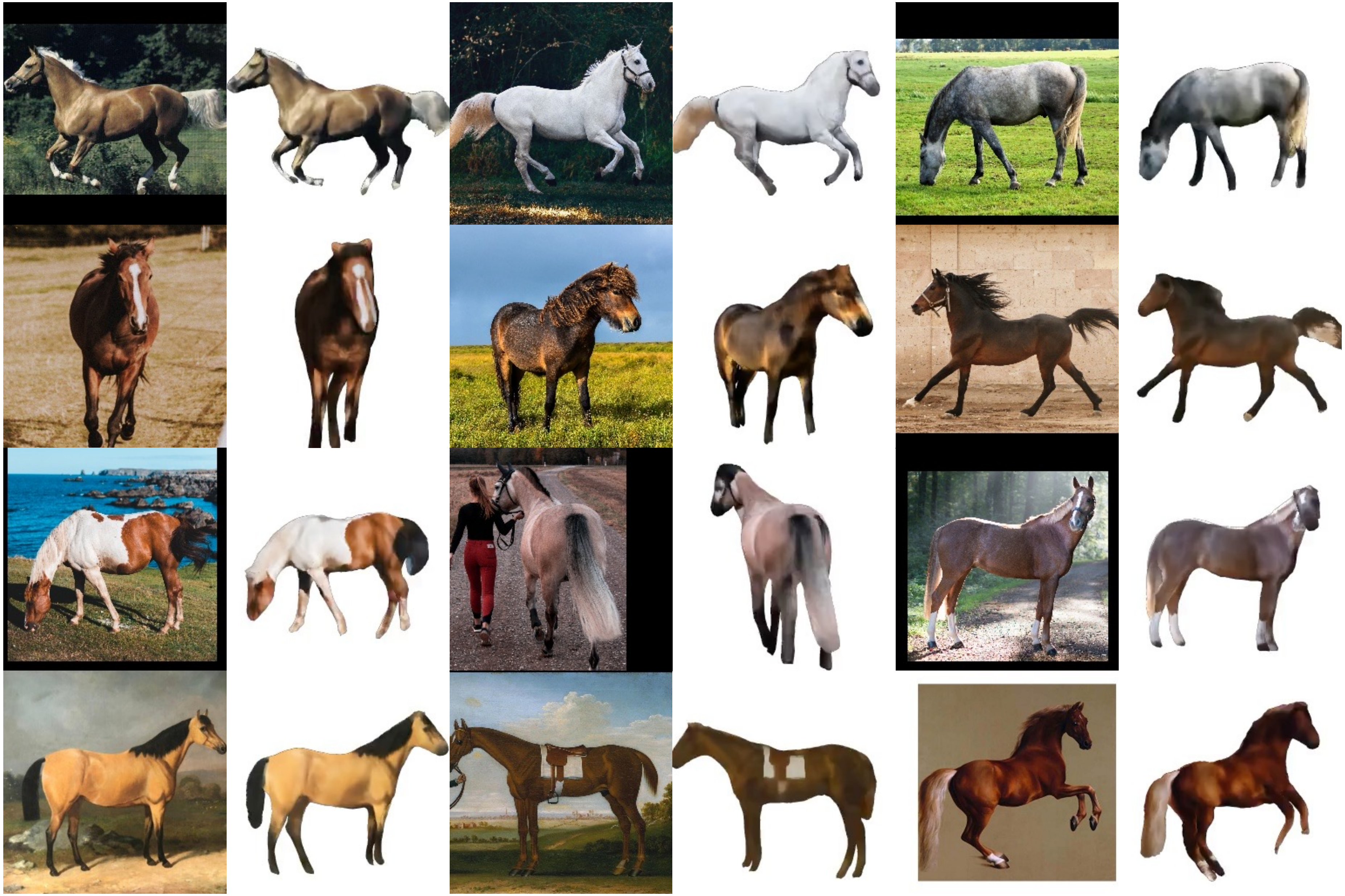
MagicPony: Learning Articulated 3D Animals in the Wild. Shangzhe Wu*, Ruining Li*, Tomas Jakab*, Christian Rupprecht, Andrea Vedaldi. CVPR 2023

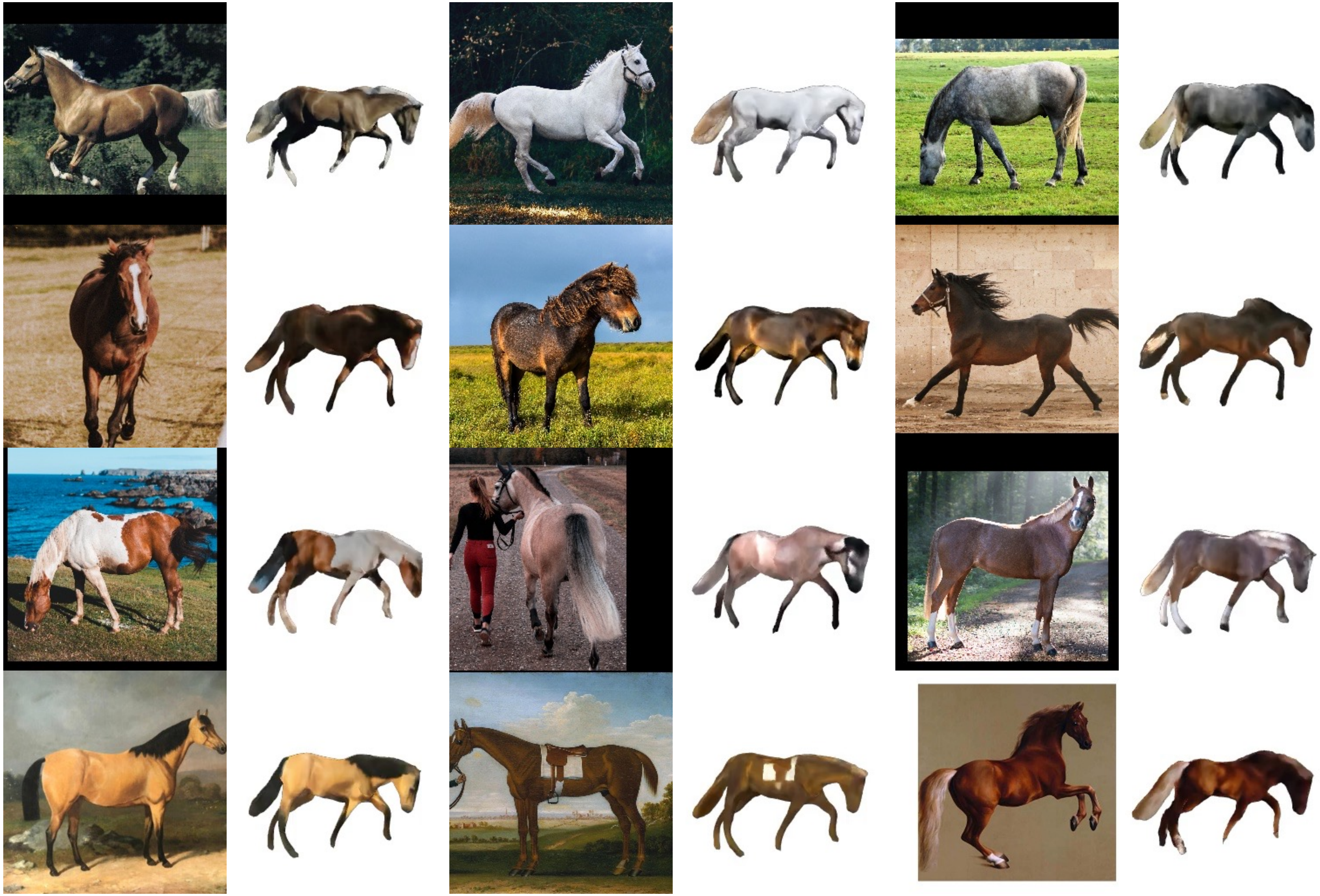
Results

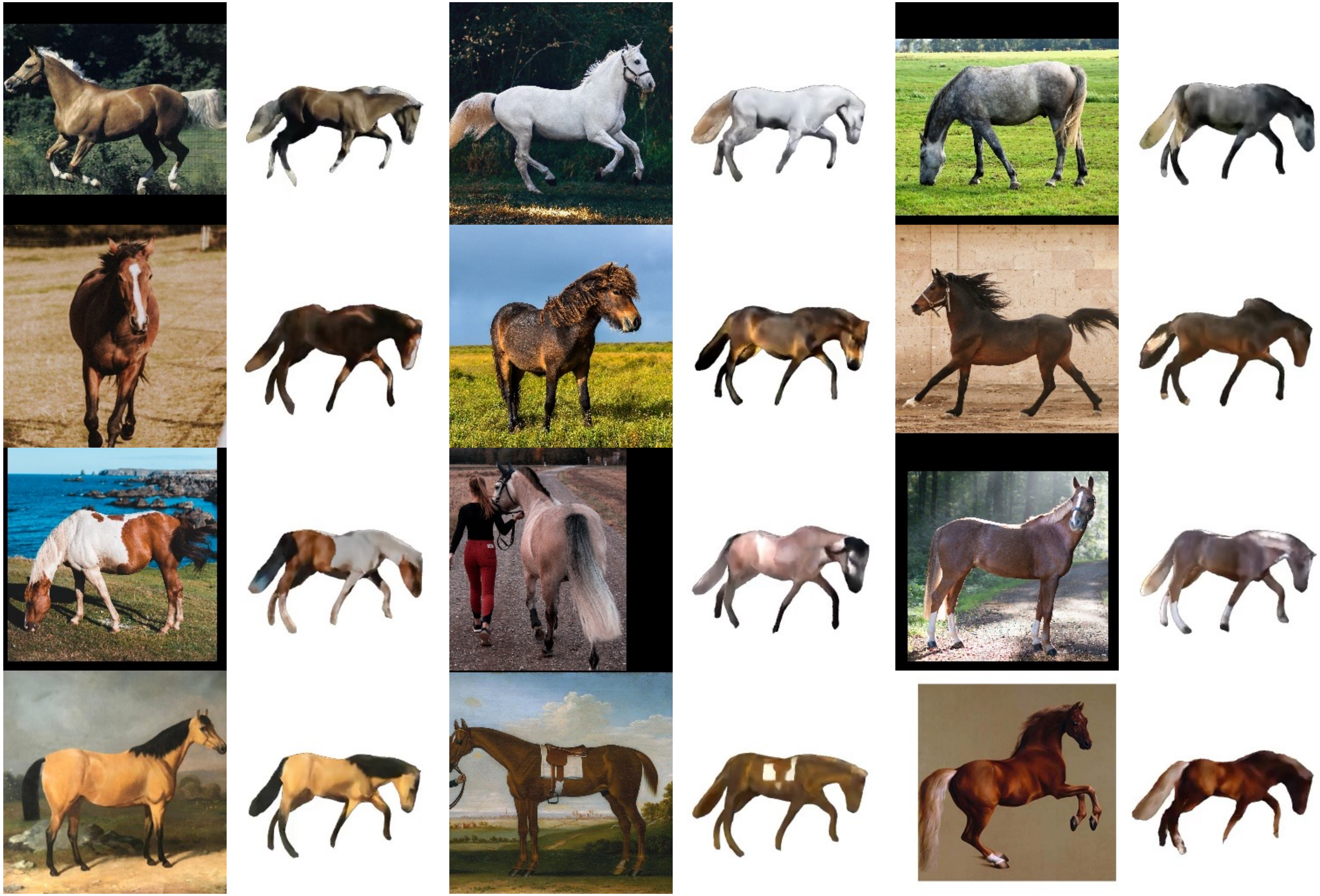


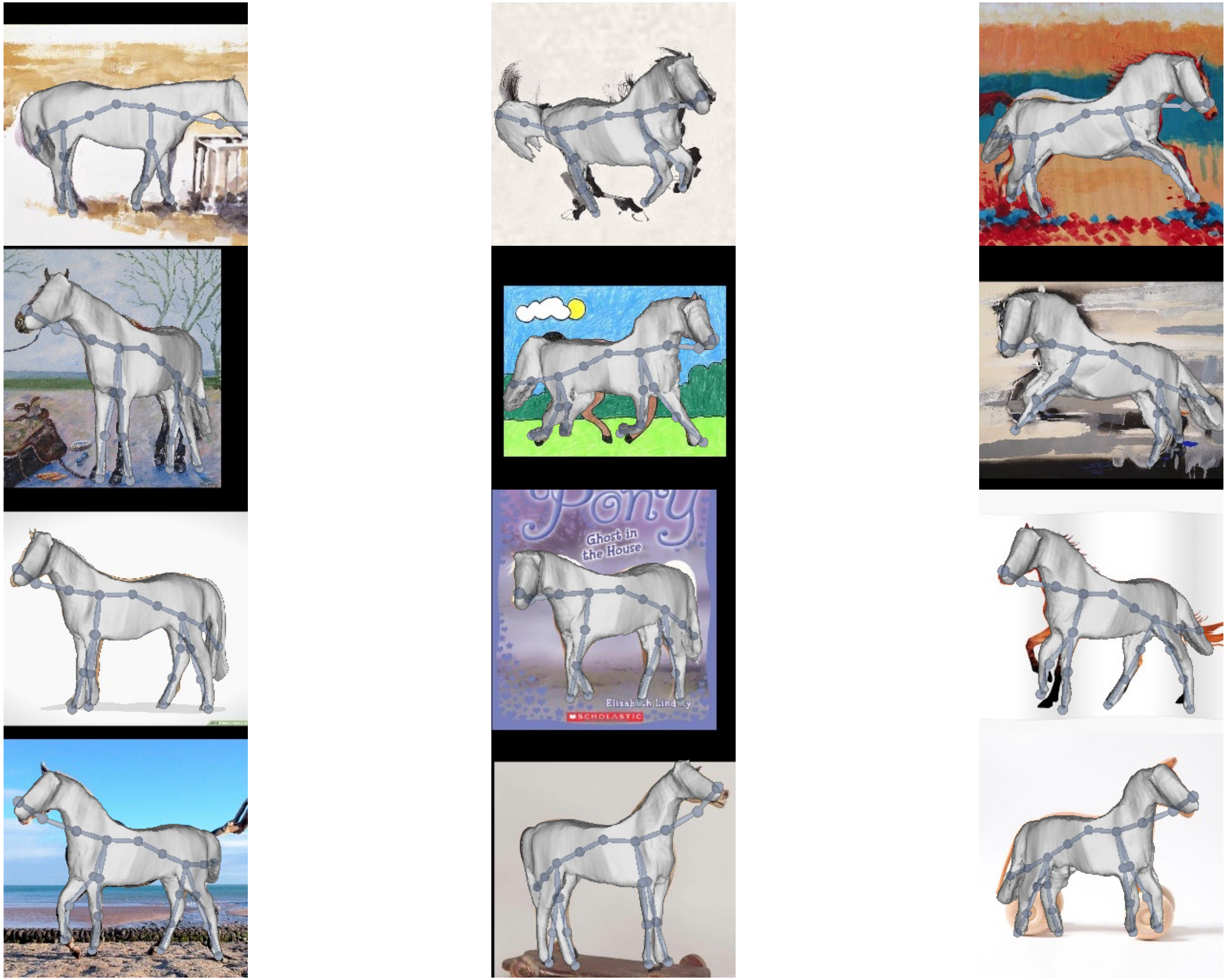


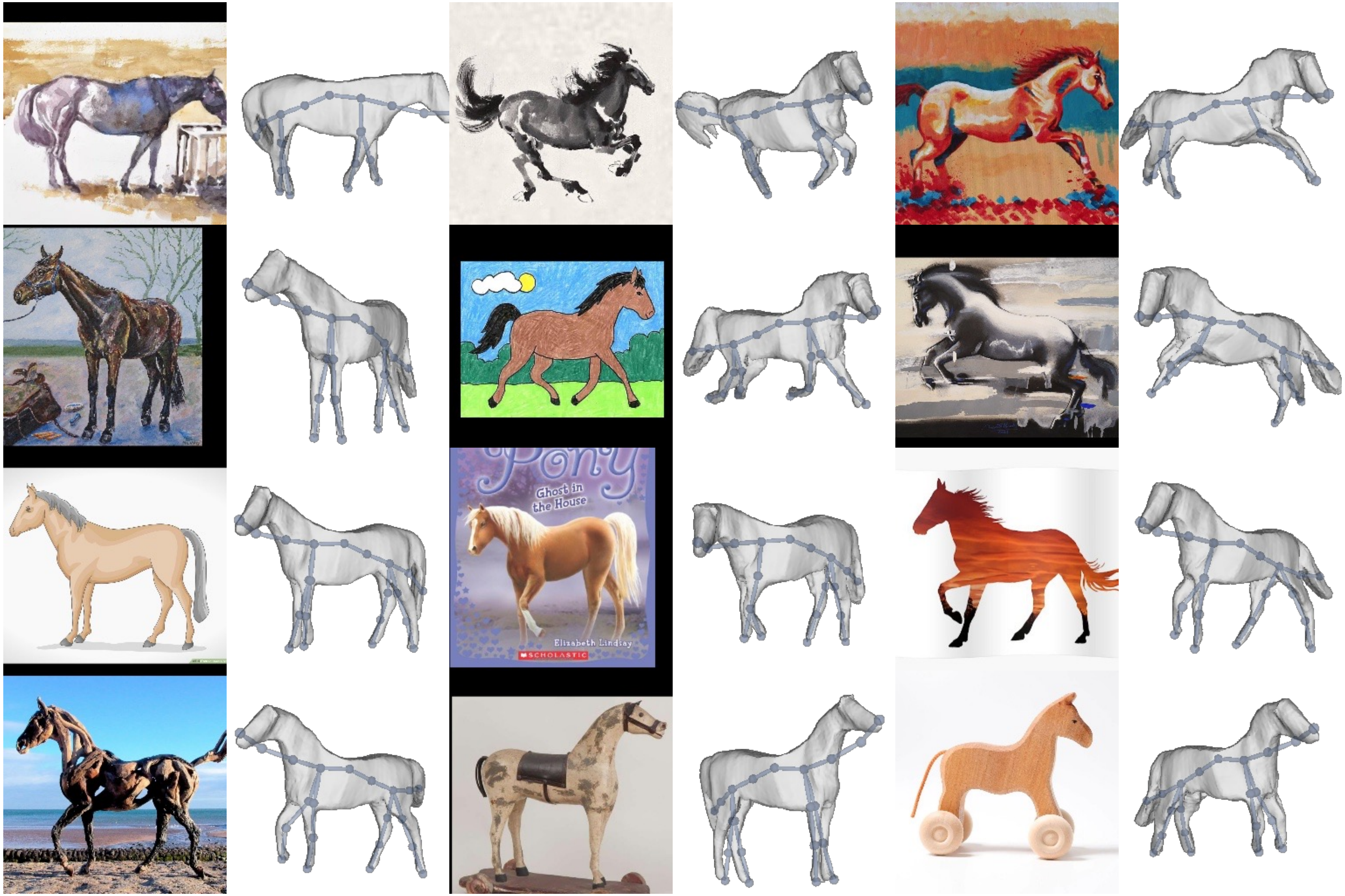


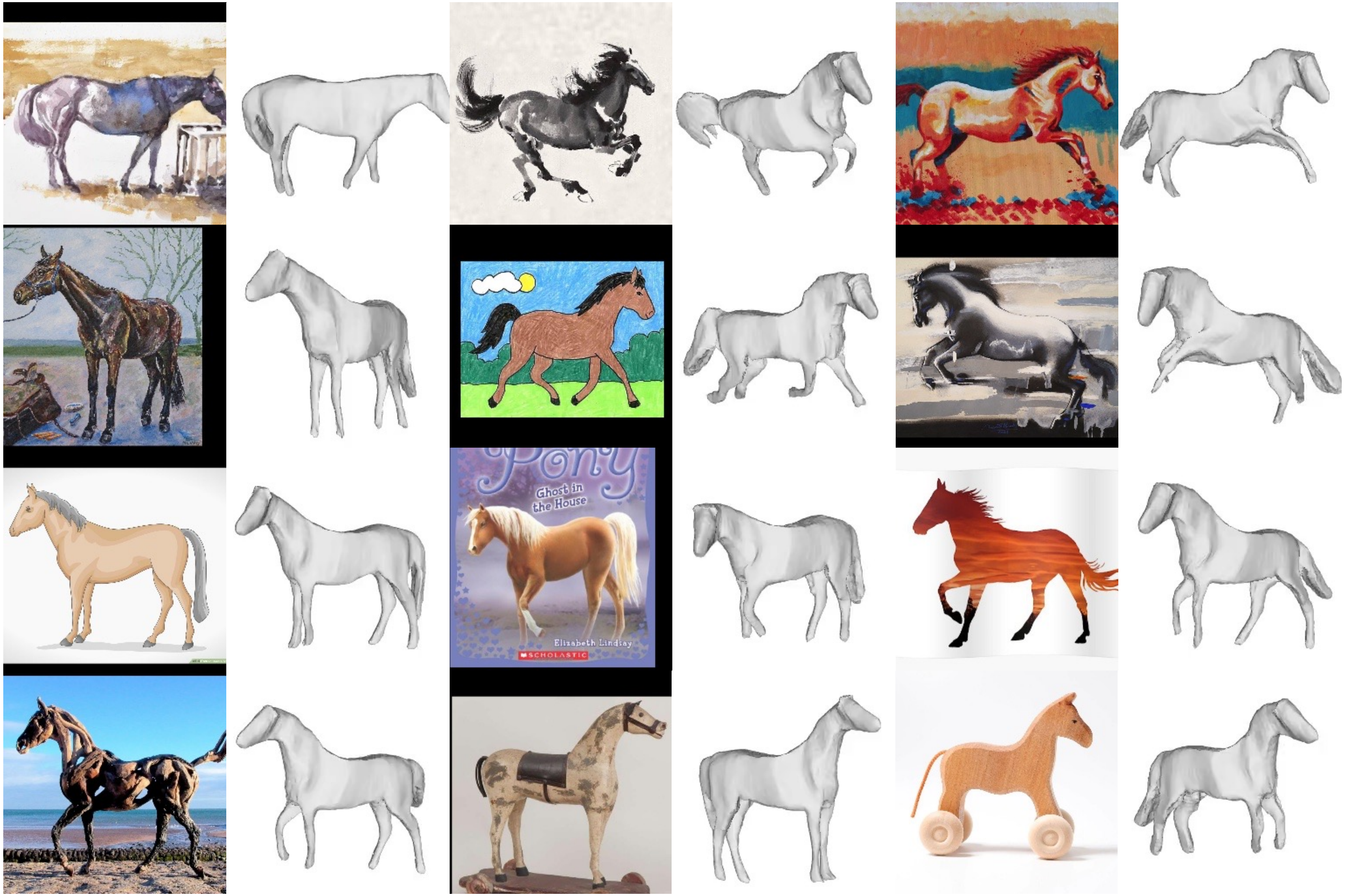


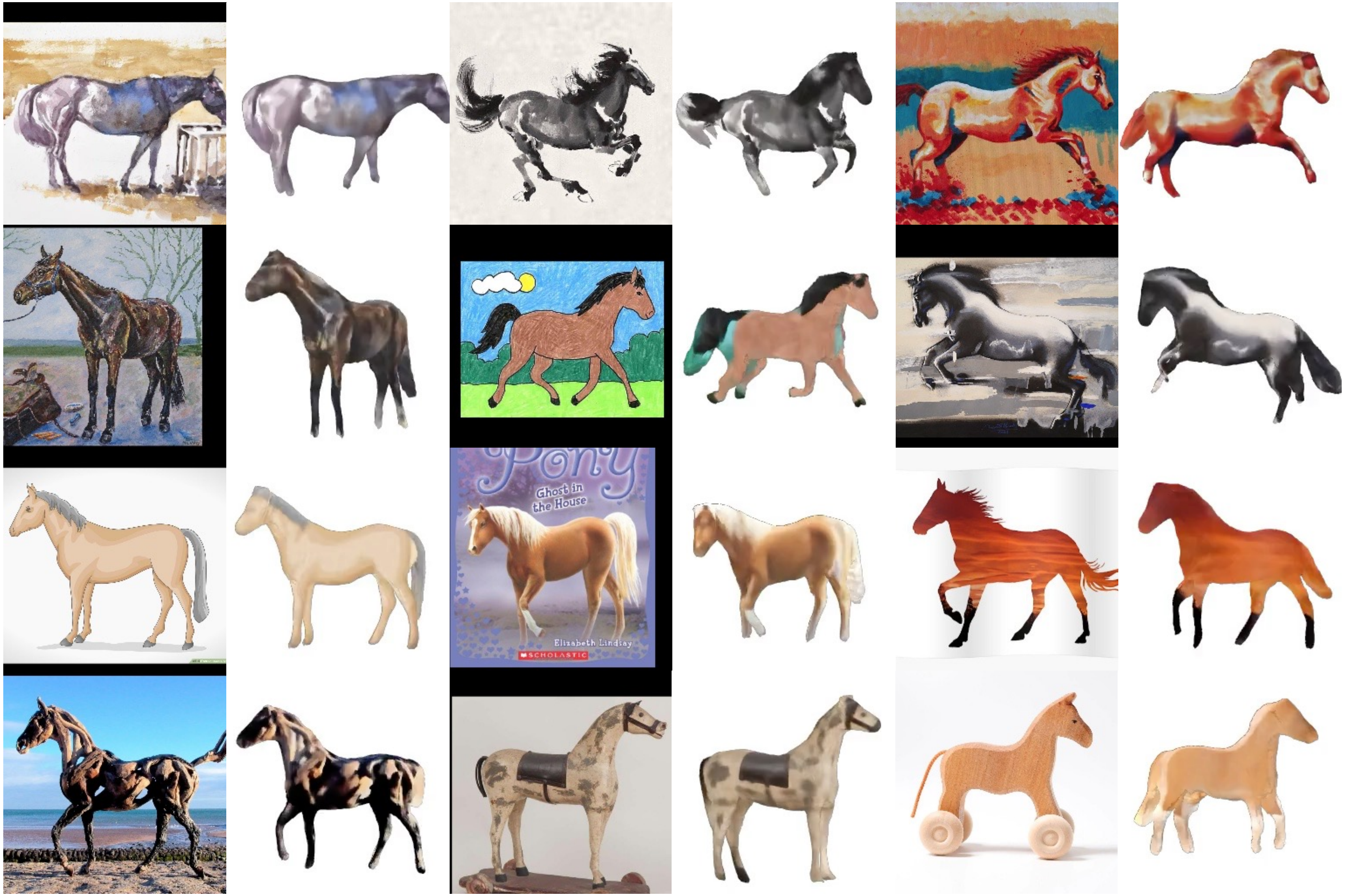


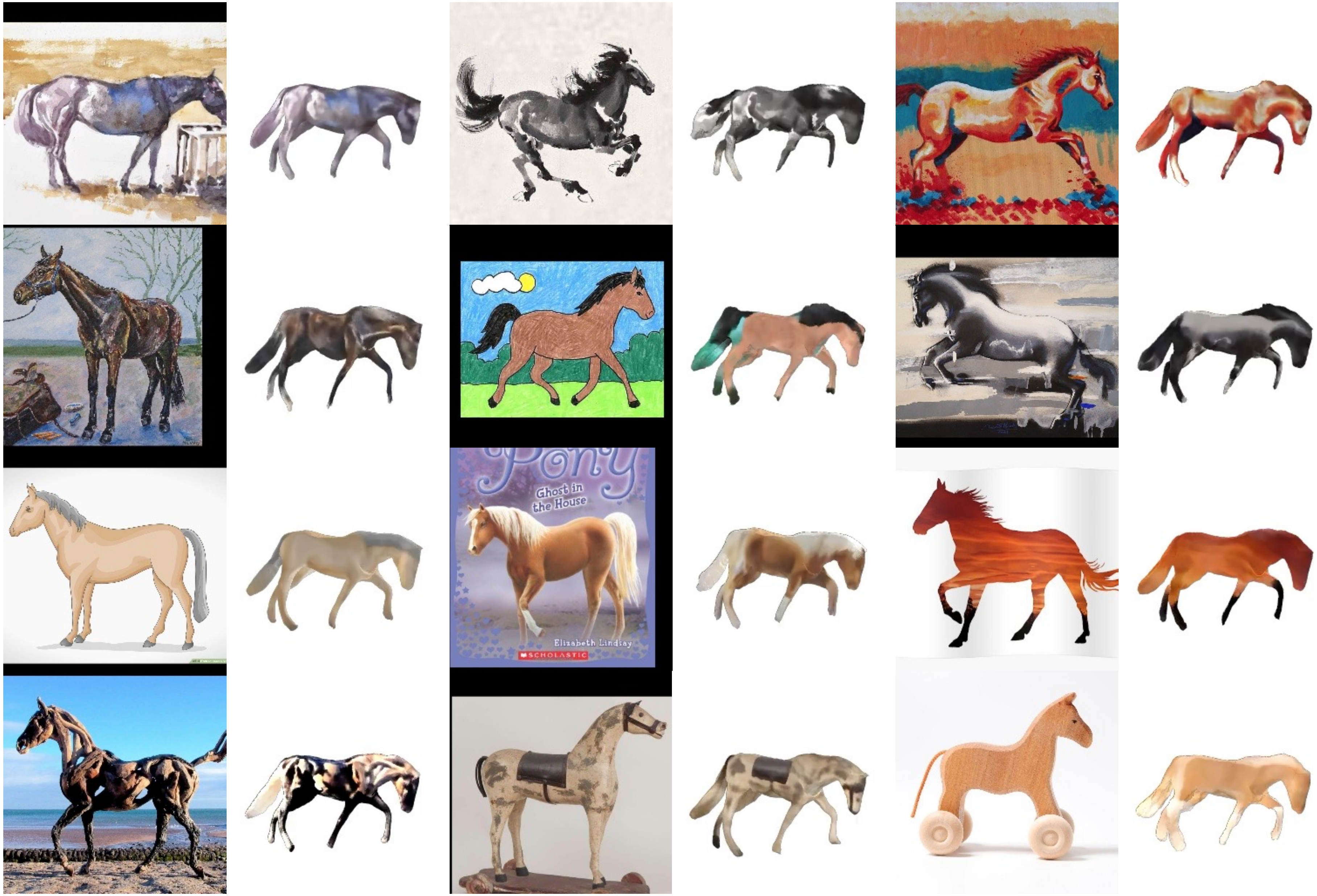


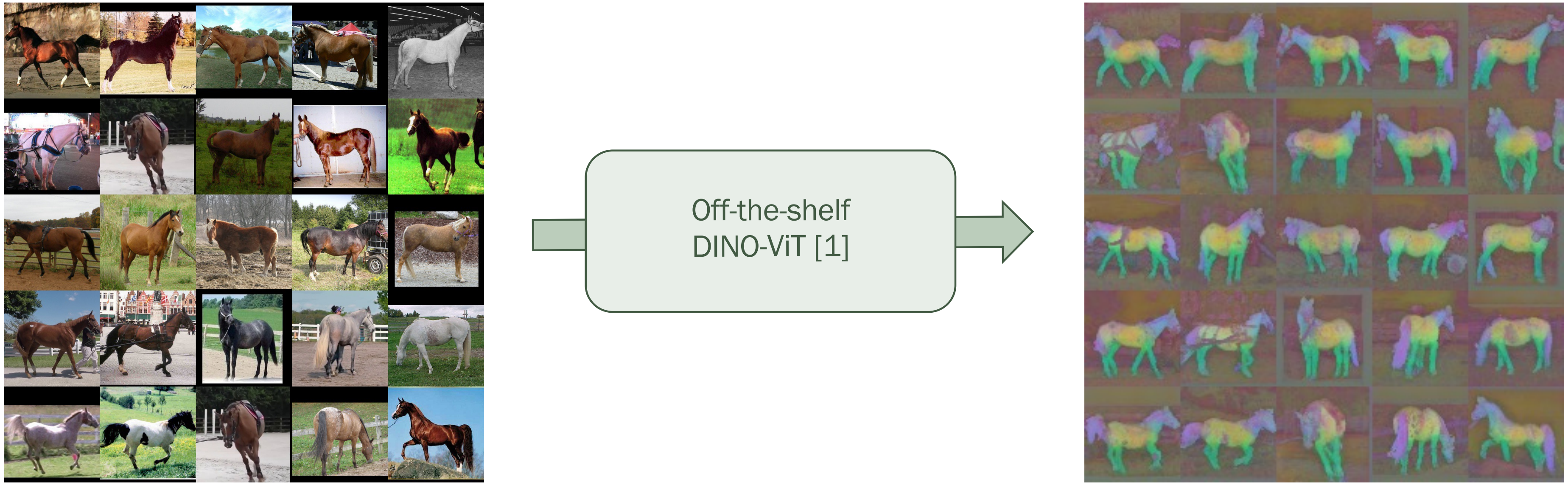

Pony
Ghost in the House
SCHOLASTIC


Pony
Ghost in the House
Elizabeth Lindsay
SCHOLASTIC


Pony
Ghost in the House
Elizabeth Lindsay
SCHOLASTIC


Pony
Ghost in the House
Elizabeth Lindsay
SCHOLASTIC


Pony
Ghost in the House
Elizabeth Lindsay
SCHOLASTIC

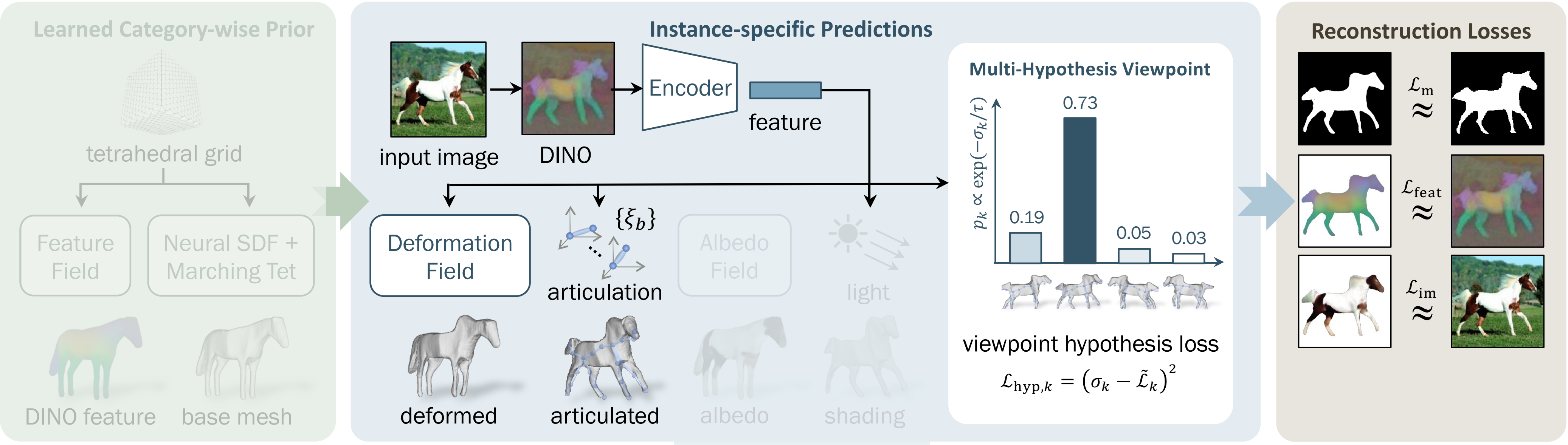


Off-the-shelf
DINO-ViT [1]

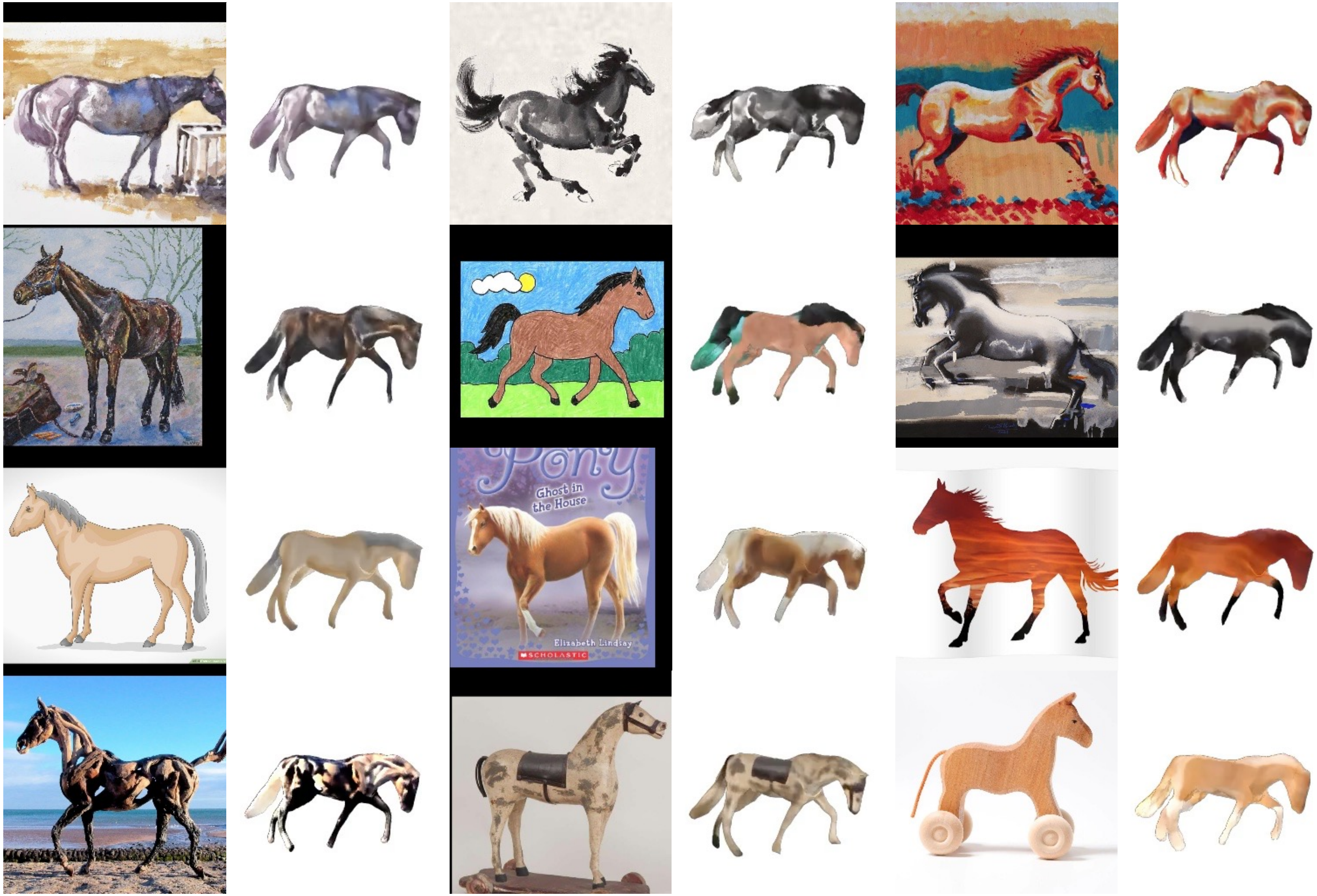
Self-supervised Image Features

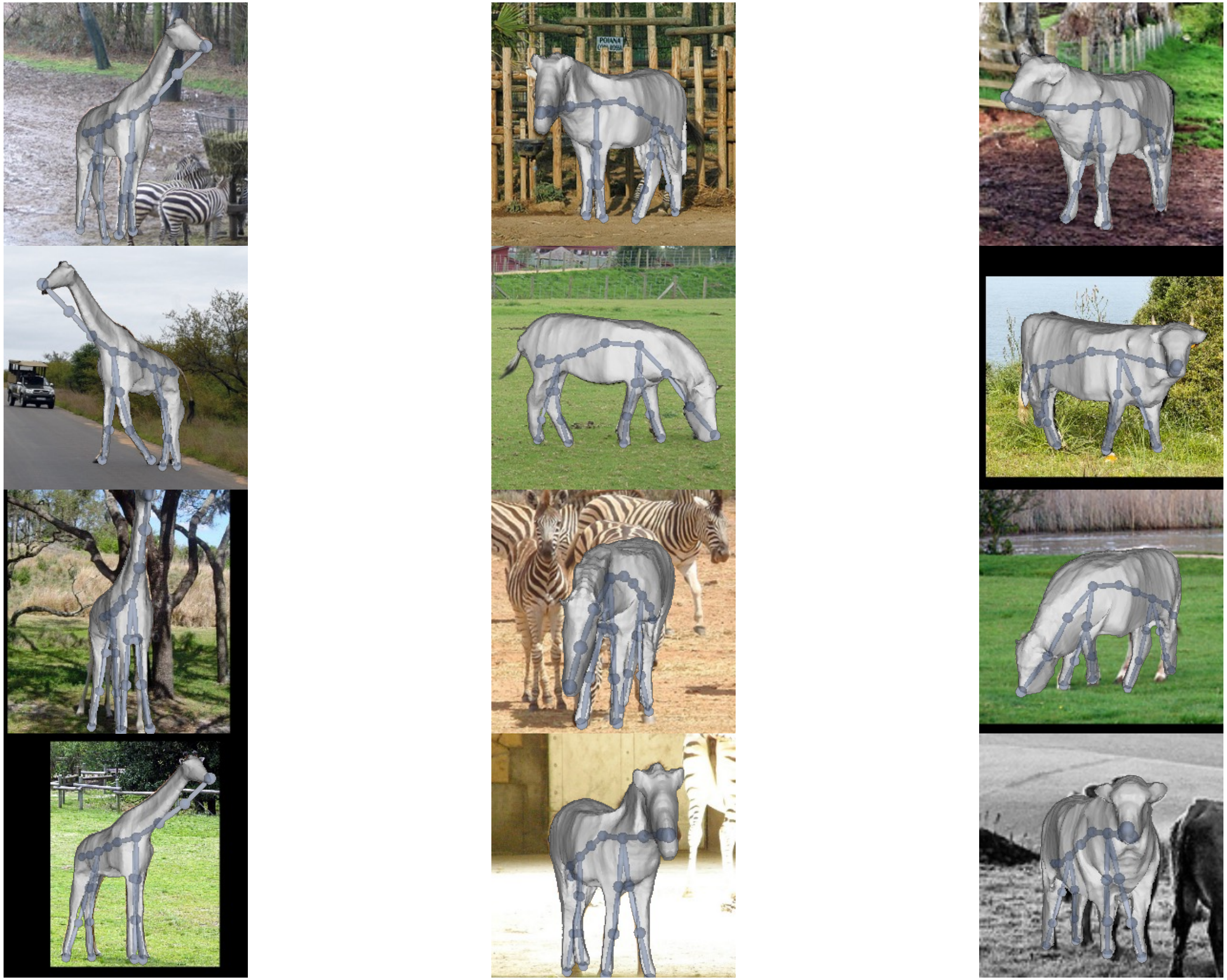
[1] Emerging Properties in Self-supervised Vision Transformers. Caron et. al. ICCV 2021.

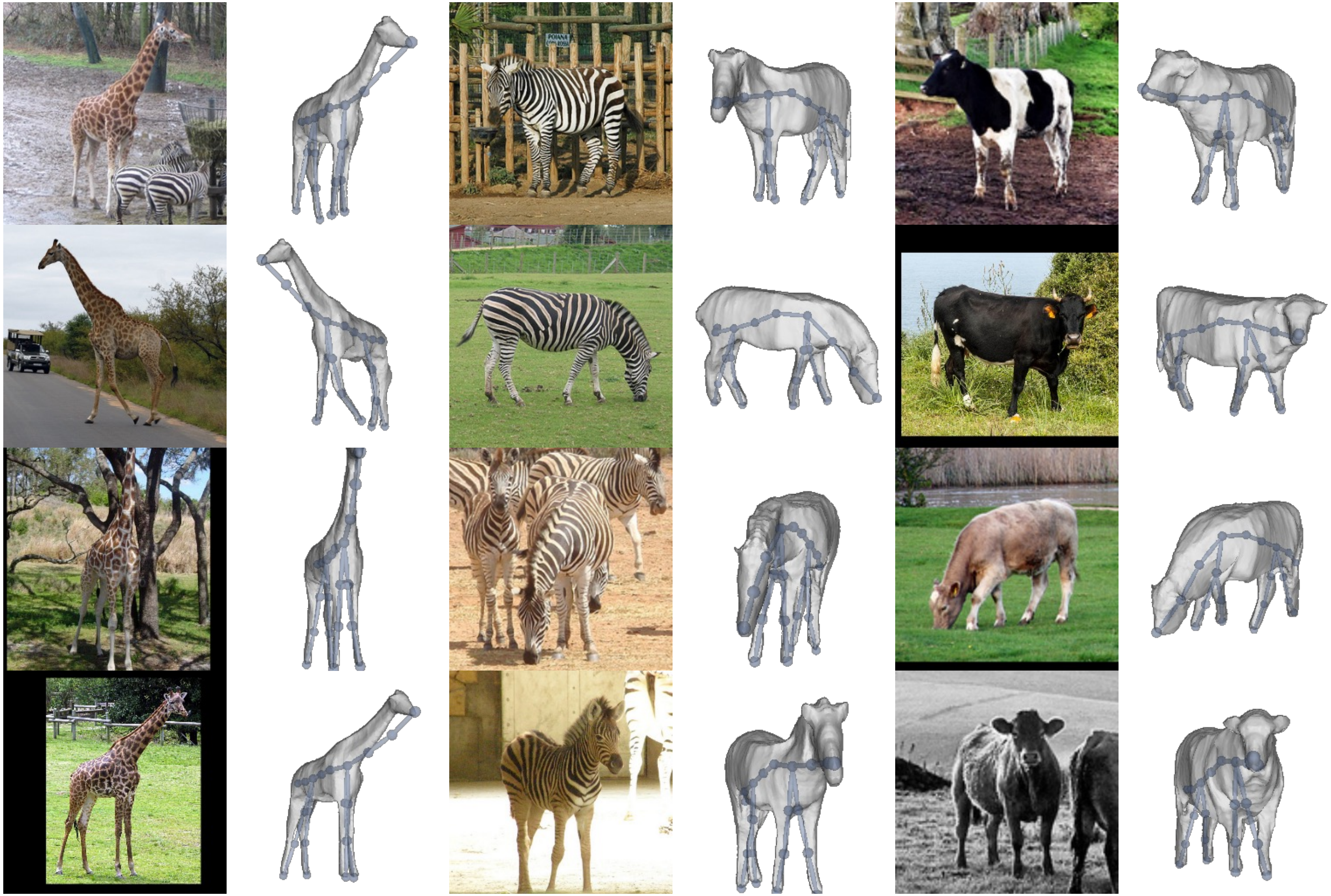
Instance Appearance

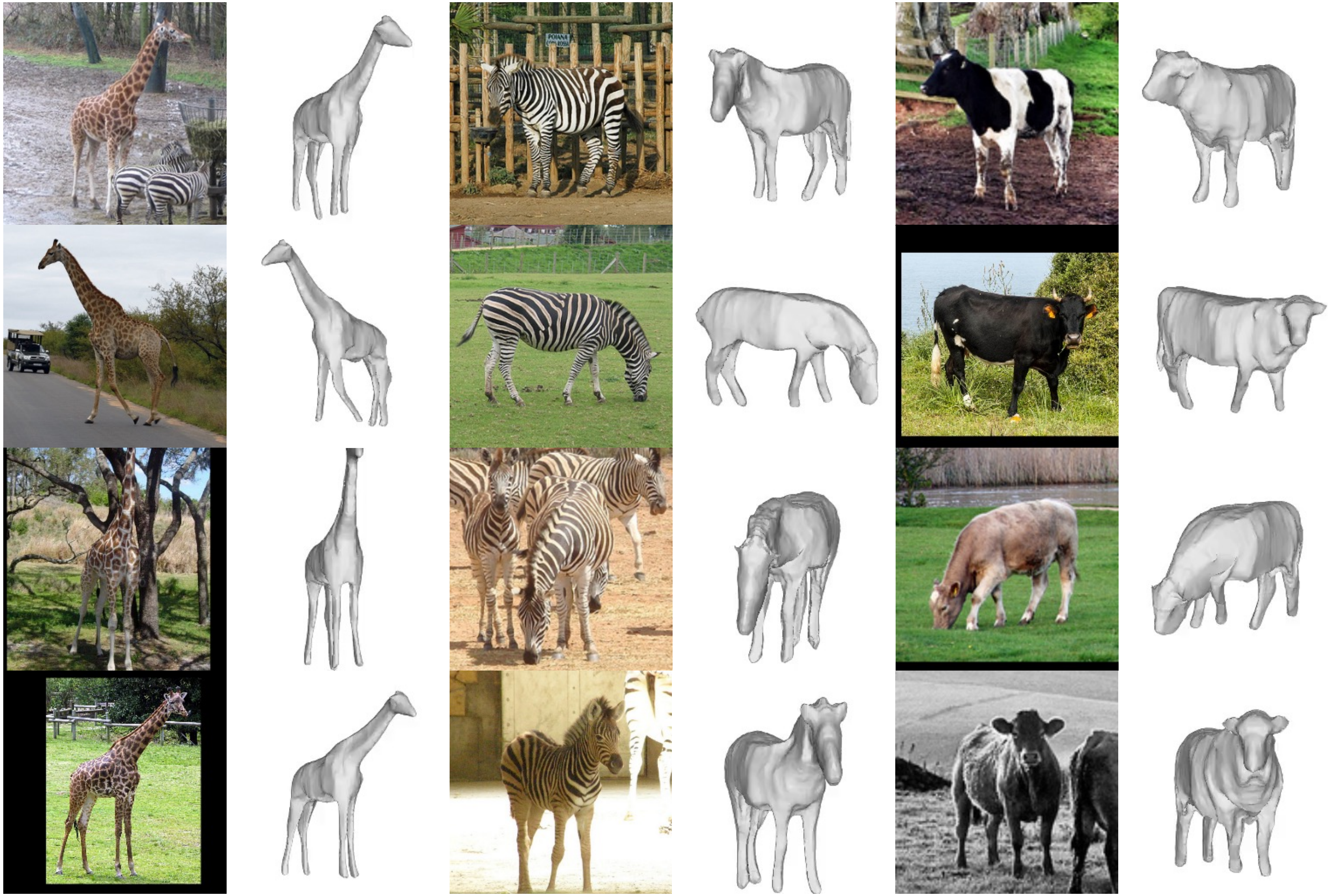


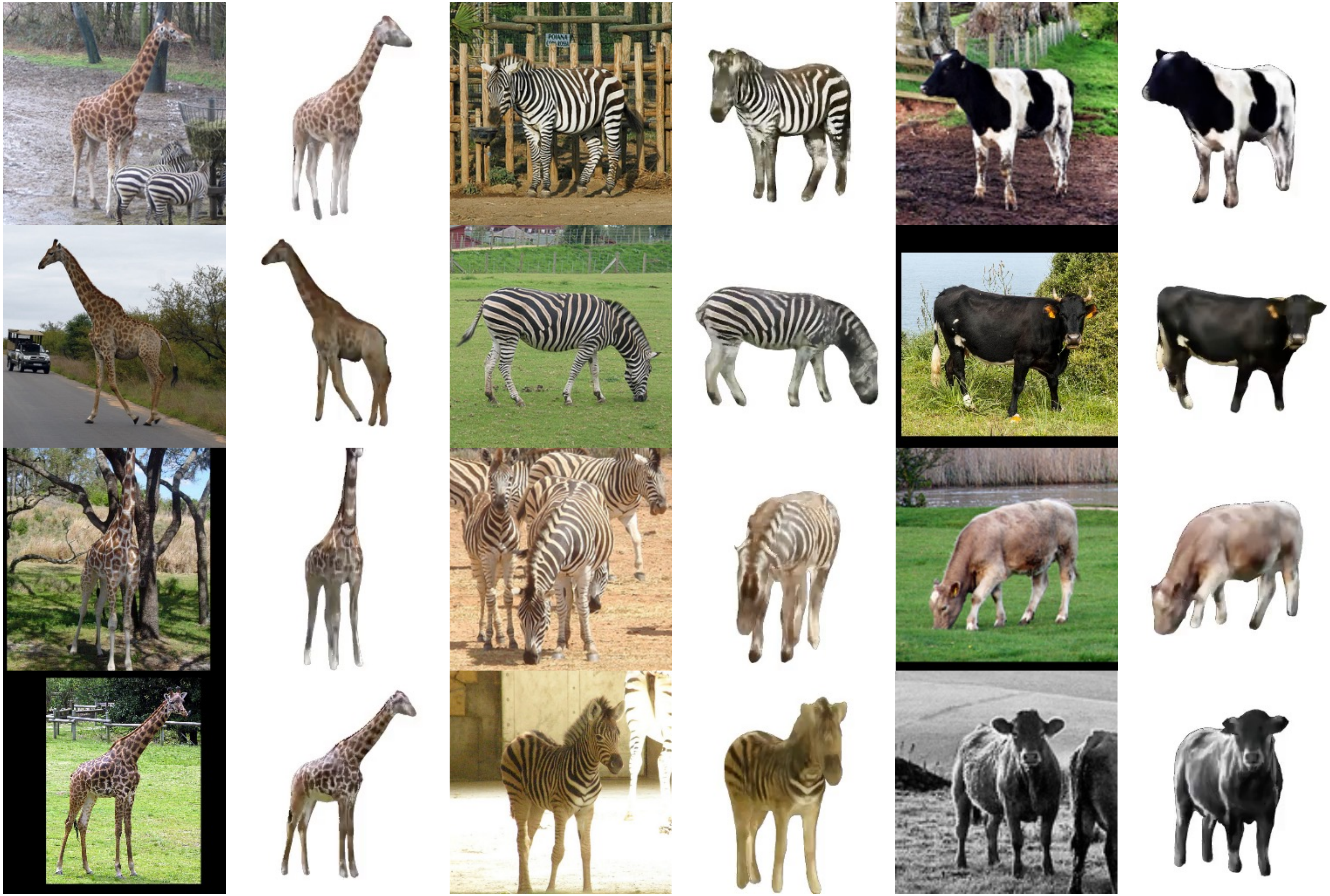
Learned Category-wise Prior
tetrahedral grid
Feature Field
Neural SDF + Marching Tet
DINO feature
base mesh
Instance-specific Predictions
input image
DINO
Encoder
feature
Deformation Field
$\{\xi_b\}$
articulation
Albedo Field
light
deformed
articulated
albedo
shading
Multi-Hypothesis Viewpoint
$p_k \propto \exp(-\sigma_k/\tau)$
0.73
0.19
0.05
0.03
viewpoint hypothesis loss
$\mathcal{L}_{\text{hyp},k} = (\sigma_k - \tilde{\mathcal{L}}_k)^2$
Reconstruction Losses
$\mathcal{L}_{\text{m}}$ ≈
$\mathcal{L}_{\text{feat}}$ ≈
$\mathcal{L}_{\text{im}}$ ≈

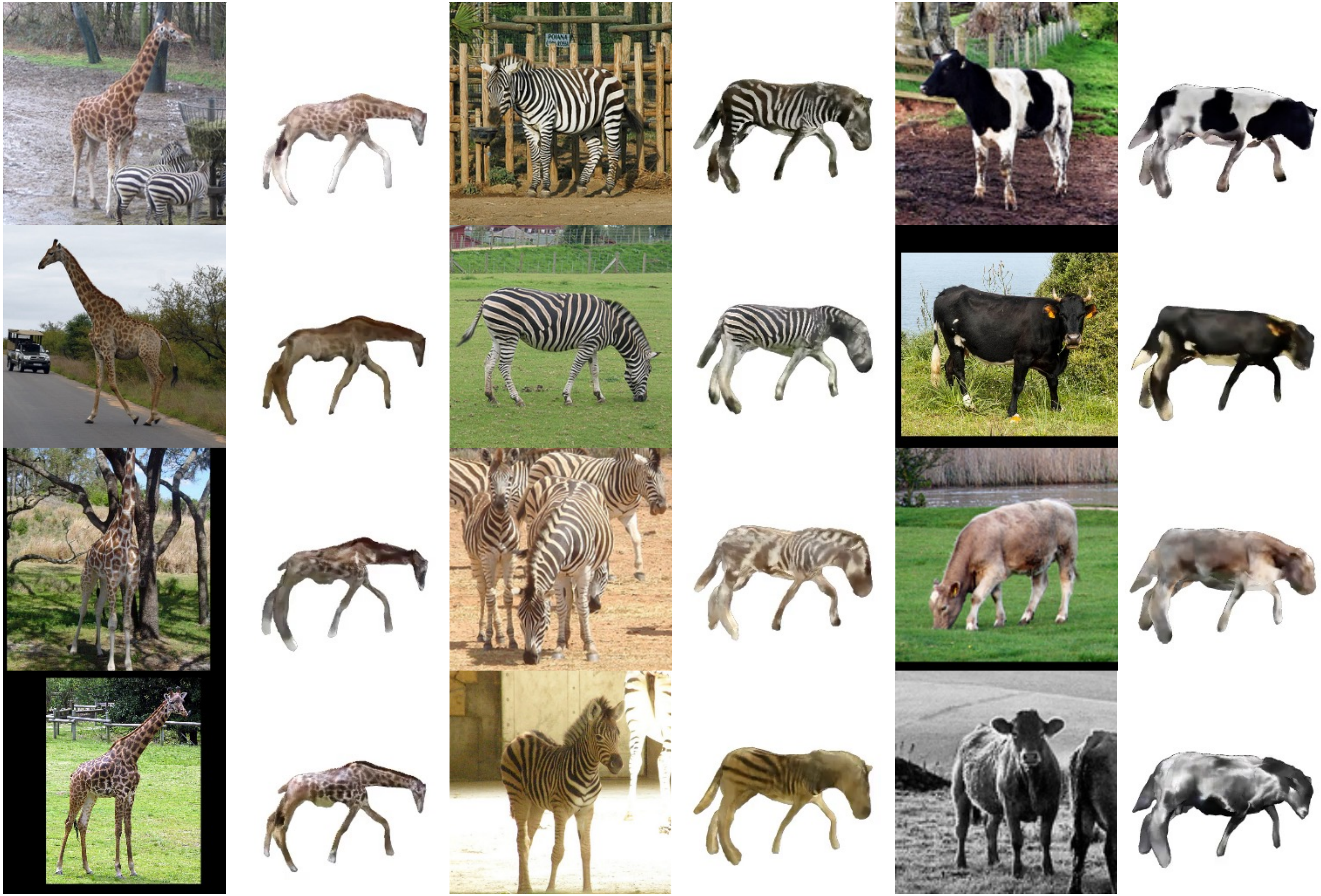

Pony
Ghost in the House
Elizabeth Lindsay
SCHOLASTIC

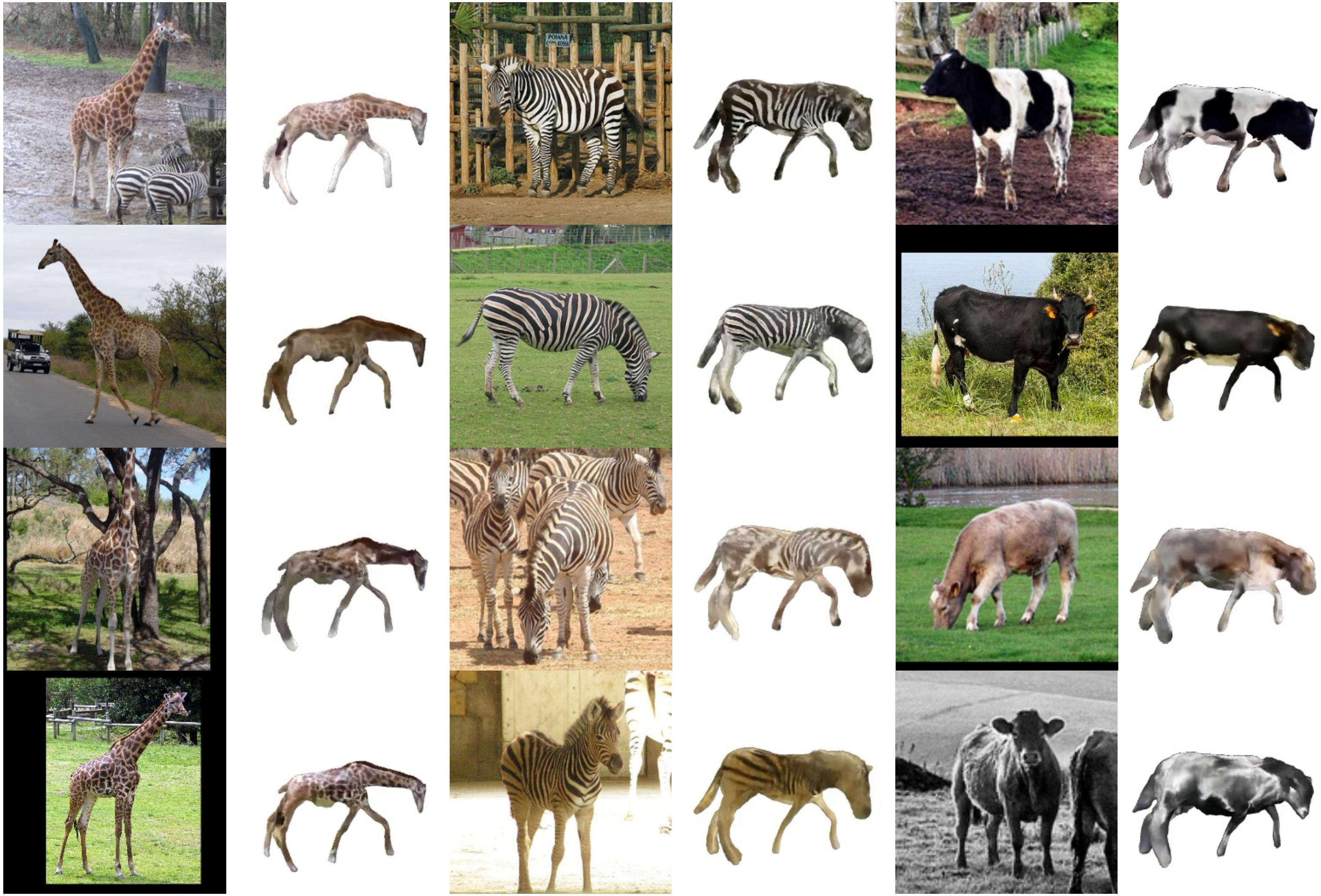




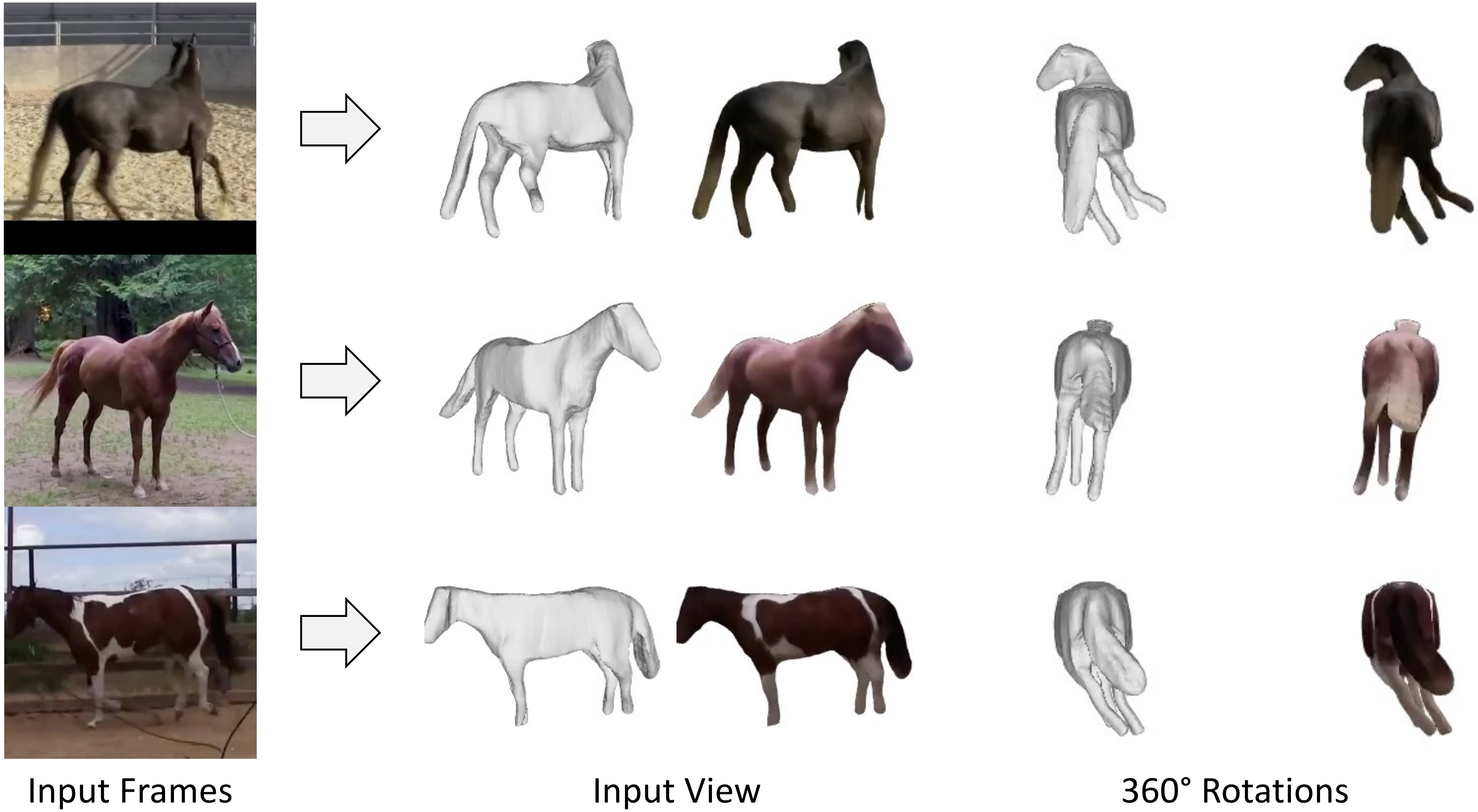








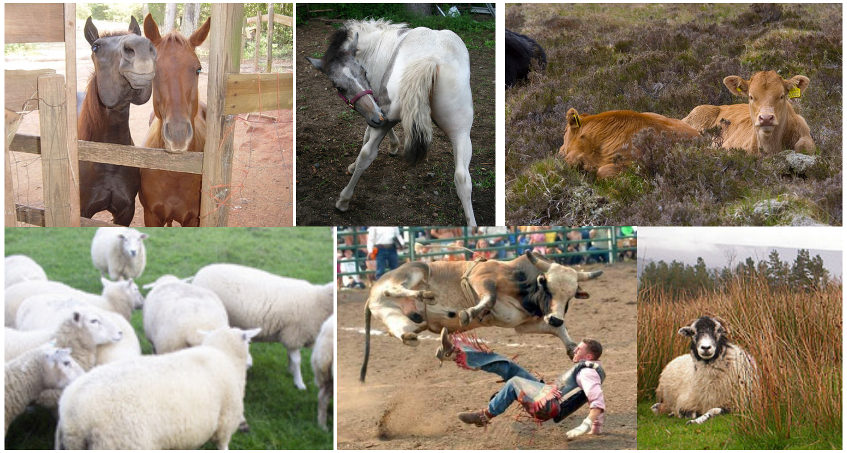
Frame-by-Frame Inference on Videos



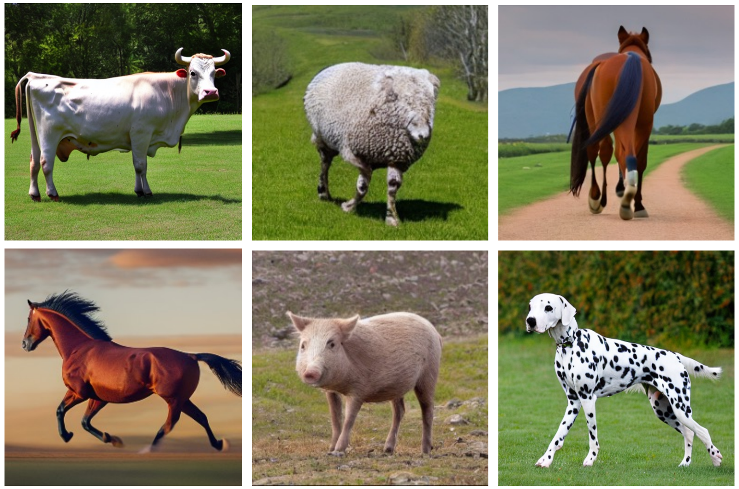
Follow up works

Real vs Diffusion Generated Images

Typical Unsuitable Real Images
from ImageNet

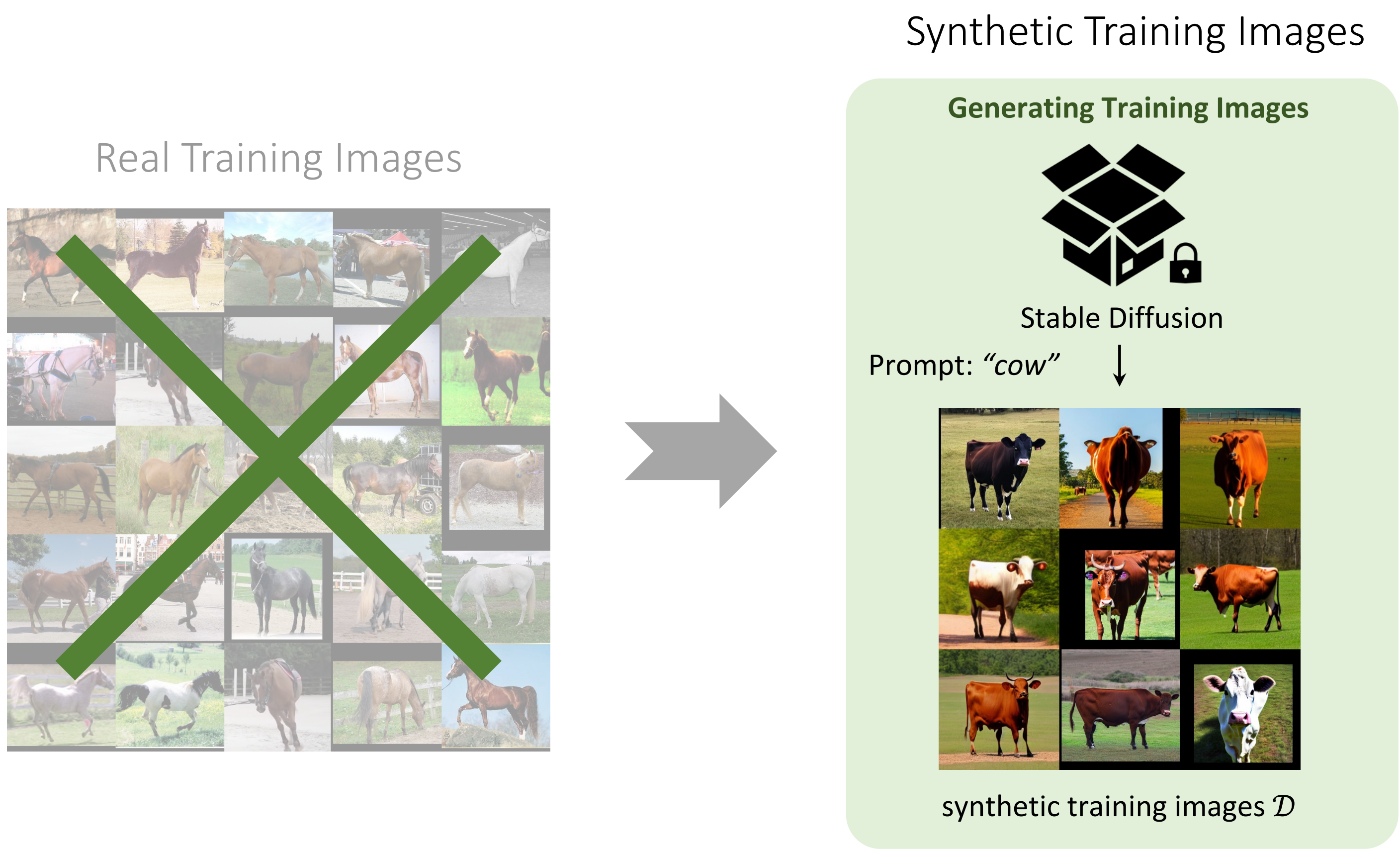


StableDiffusion Generated Images



“Implicitly curated”

Synthetic Training Images



Real Training Images

Synthetic Training Images

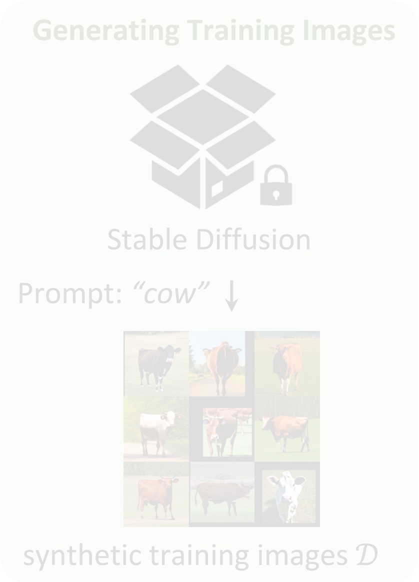
Generating Training Images

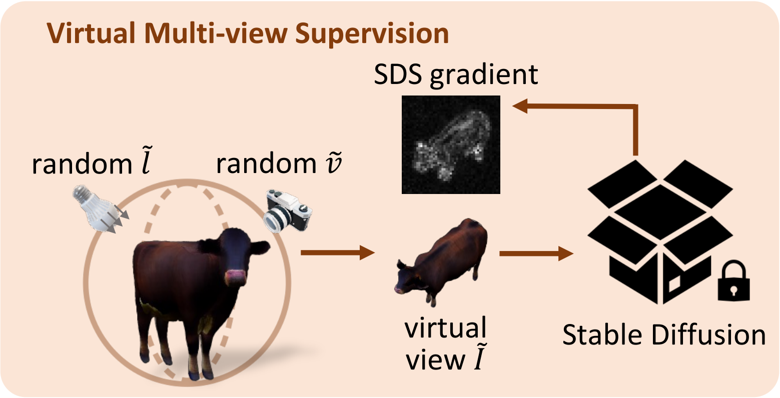
Stable Diffusion

Prompt: *“cow”*

synthetic training images $\mathcal{D}$

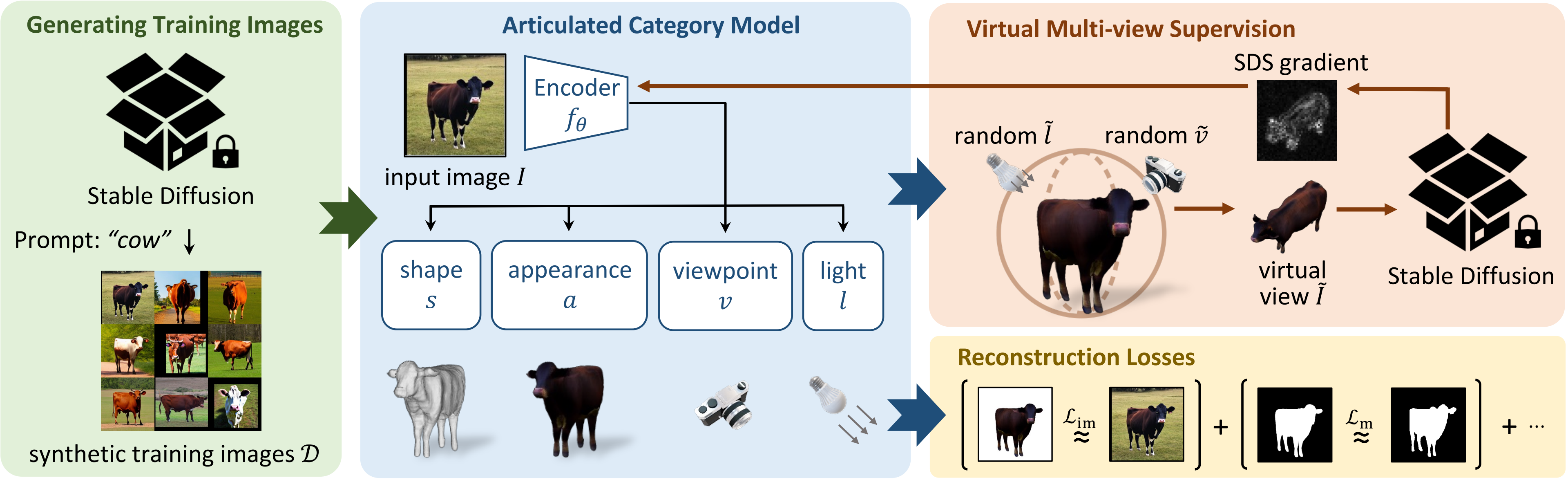
Virtual Multi-view Supervision


Generating Training Images
Stable Diffusion
Prompt: *"cow"*
synthetic training images $\mathcal{D}$

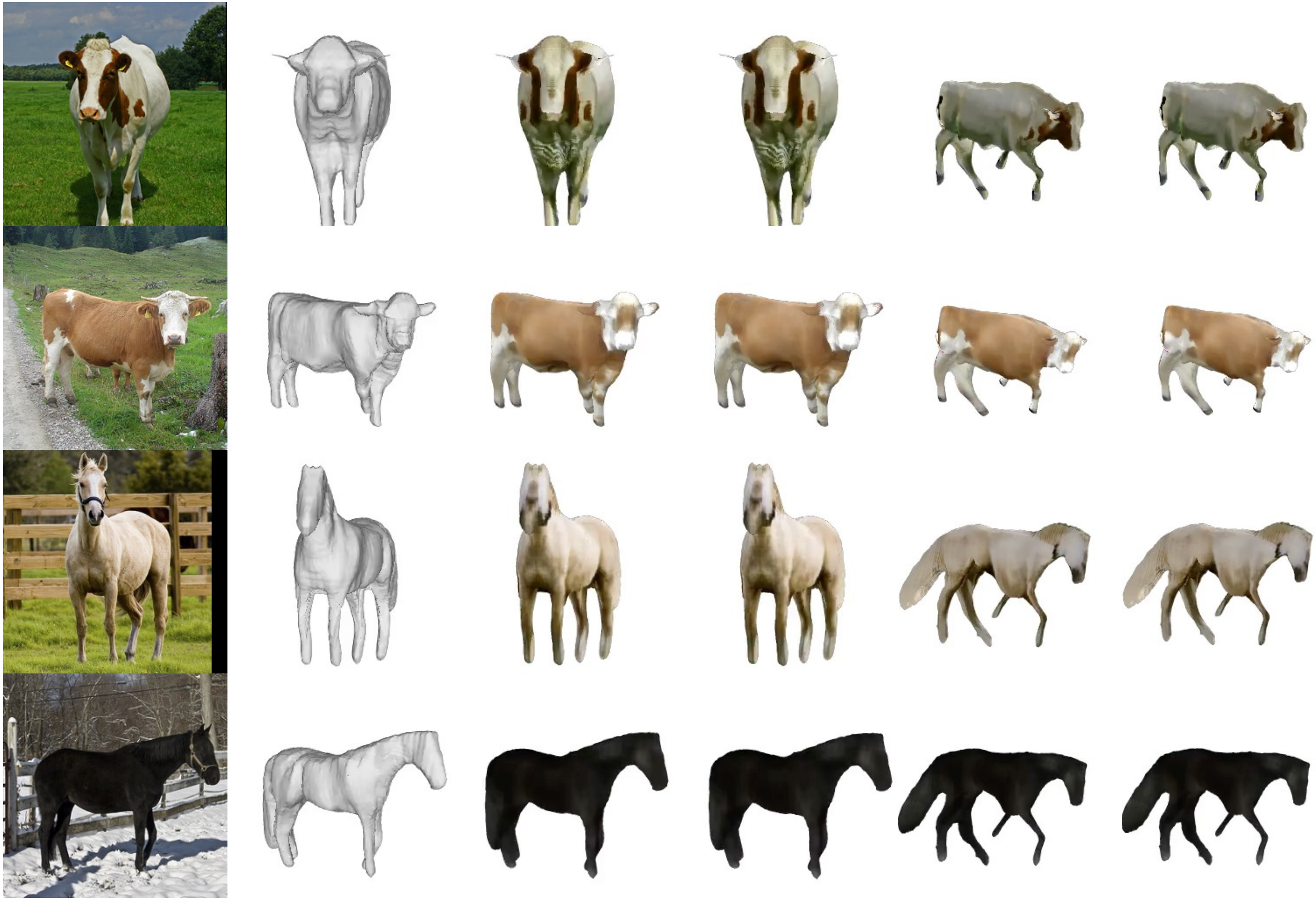


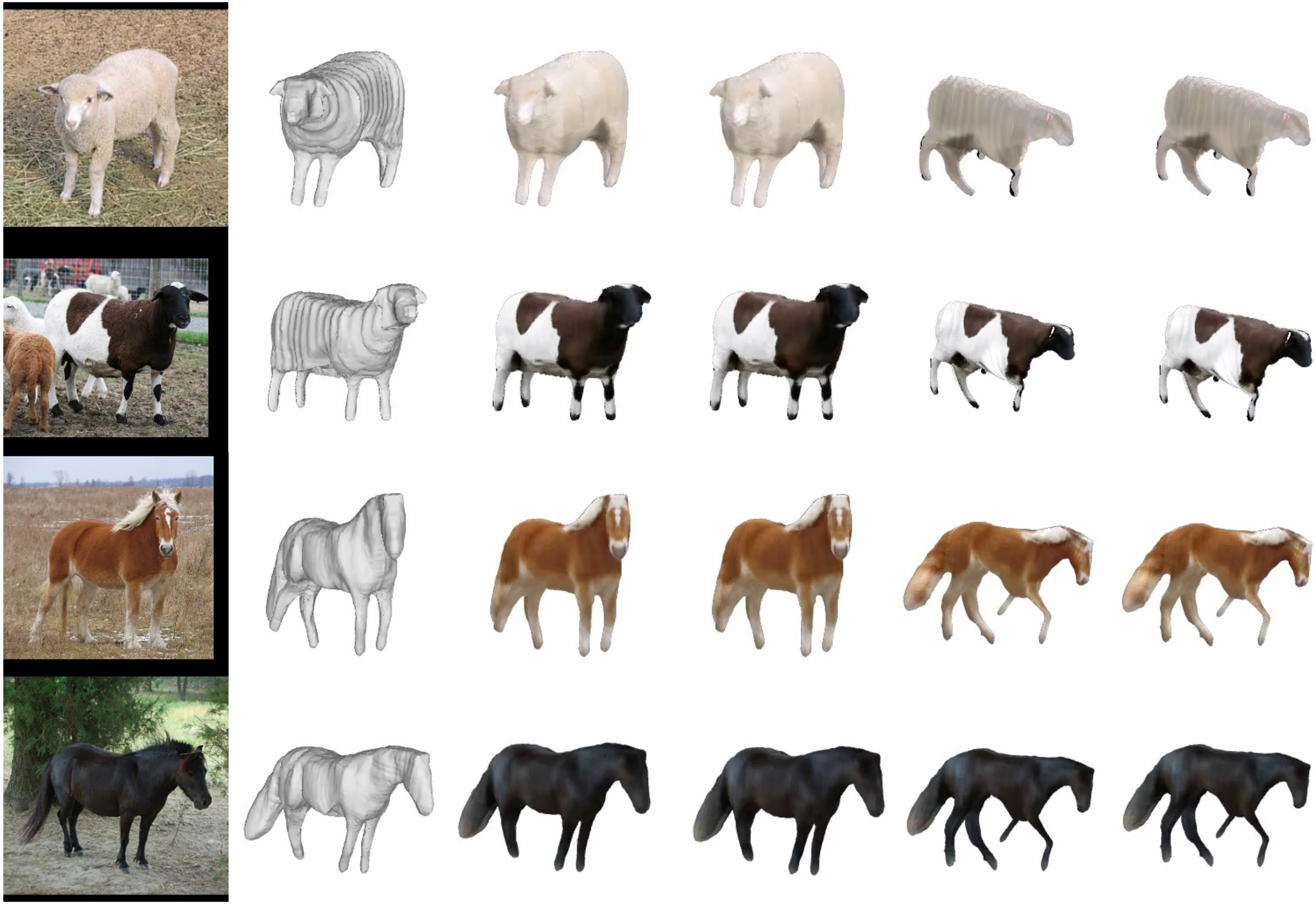
[1] DreamFusion: Text-to-3D using 2D Diffusion. Poole et. al. arXiv 2022.

Training pipeline

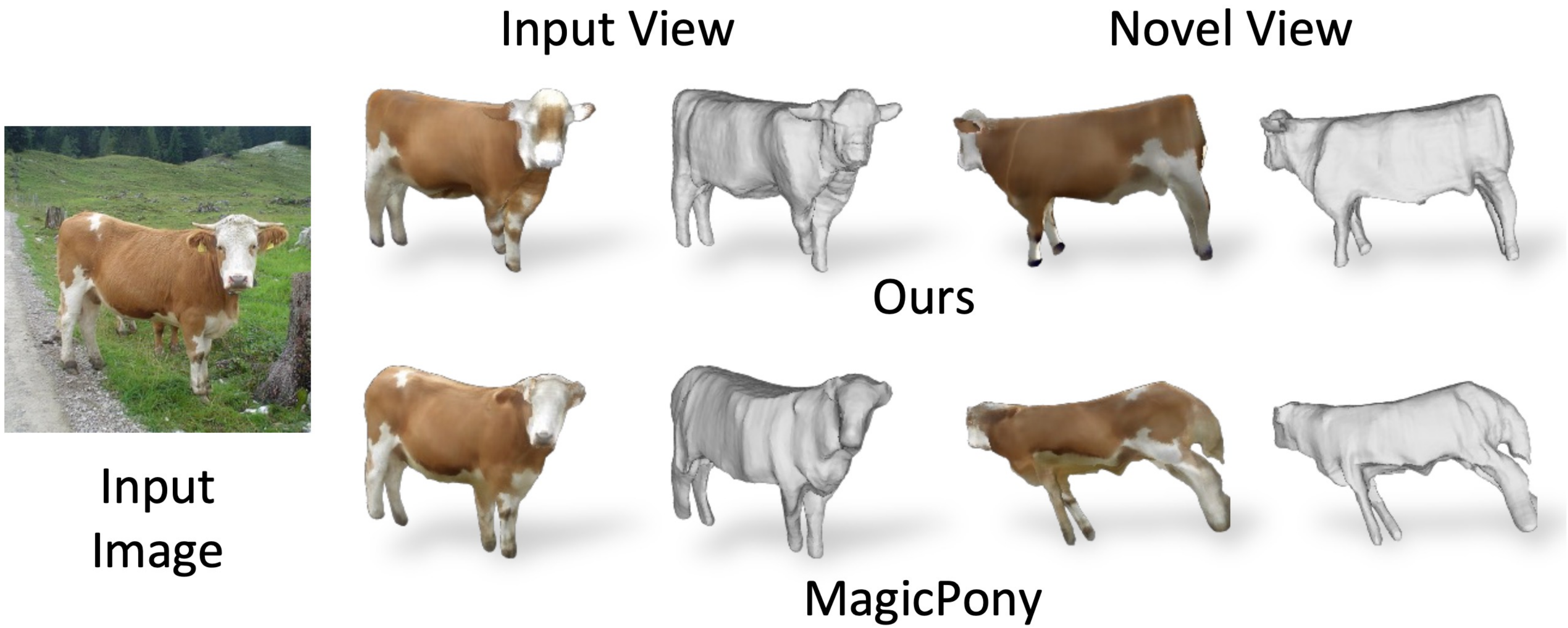


Farm3D: Learning Articulated 3D Animals by Distilling 2D Diffusion. Tomas Jakab*, Ruining Li*, Shangzhe Wu, Christian Rupprecht, Andrea Vedaldi. 3DV 2024





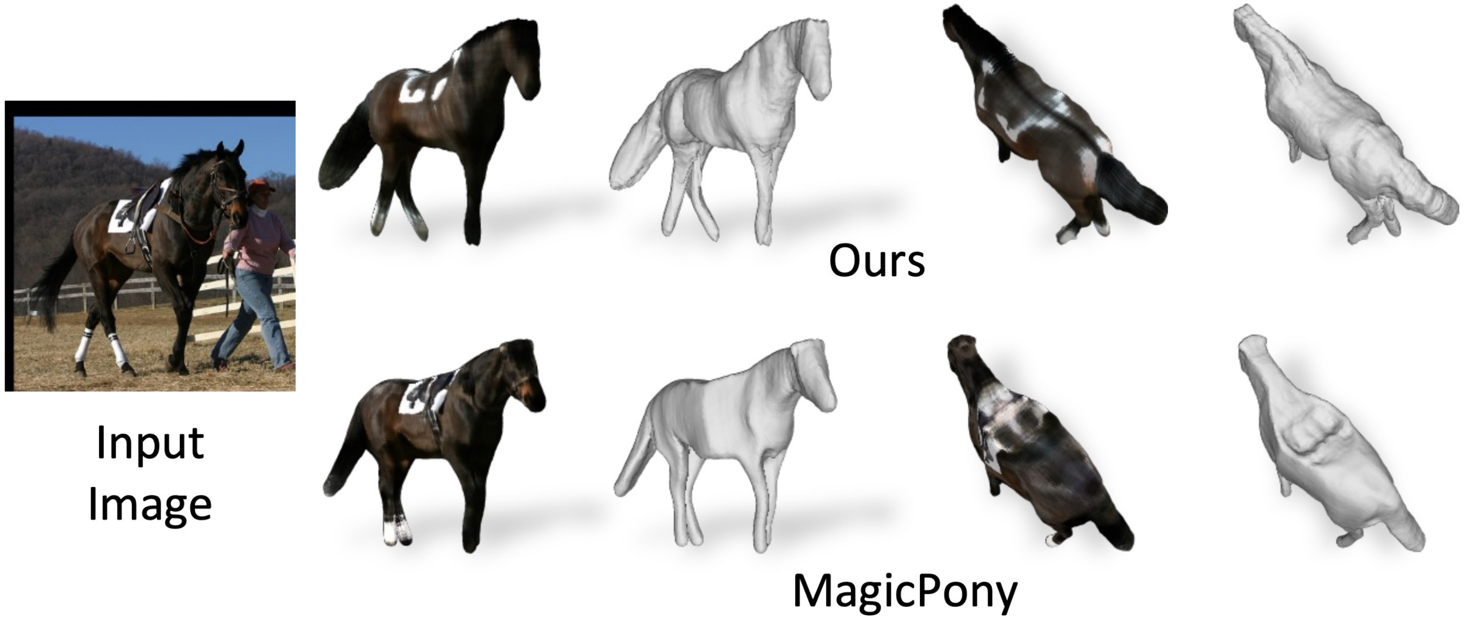
Comparison with MagicPony



MagicPony: Learning Articulated 3D Animals in the Wild

Shangzhe Wu, Ruining Li*, Tomas Jakab*, Christian Rupprecht, Andrea Vedaldi*, CVPR 2023, **equal contribution*

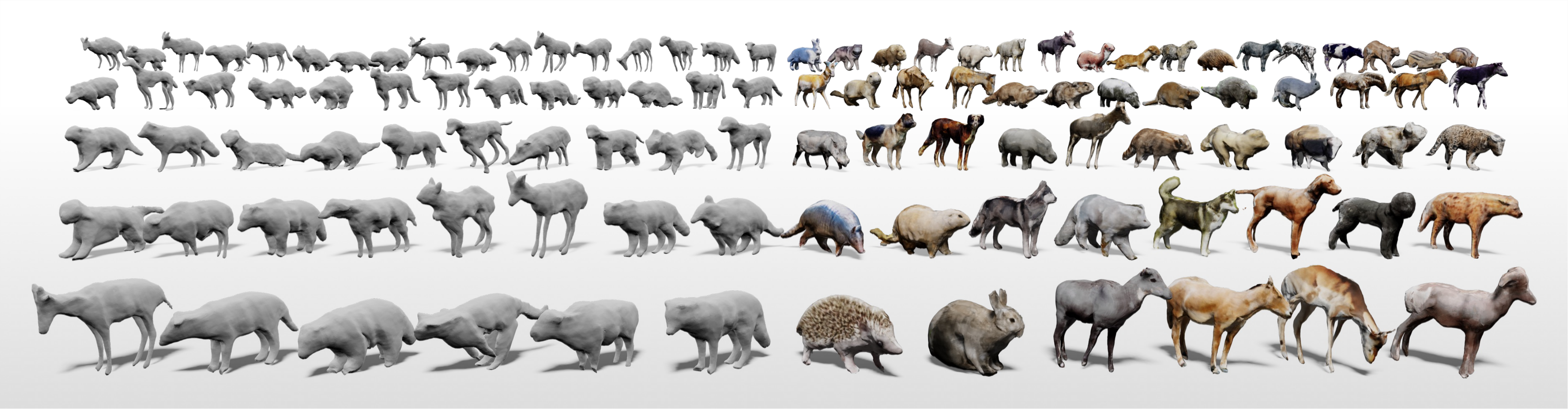
Comparison with MagicPony



MagicPony: Learning Articulated 3D Animals in the Wild

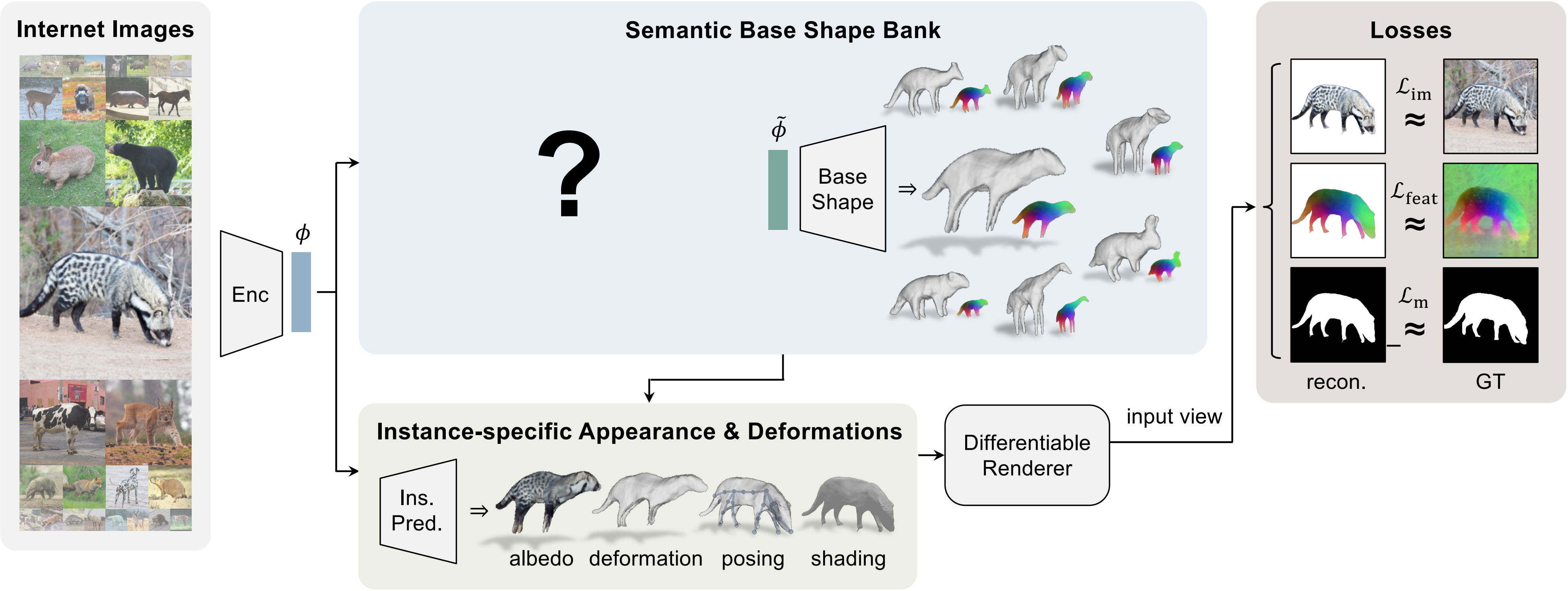
Shangzhe Wu, Ruining Li*, Tomas Jakab*, Christian Rupprecht, Andrea Vedaldi*, CVPR 2023, **equal contribution*

Towards Reconstructing the Animal Kingdom

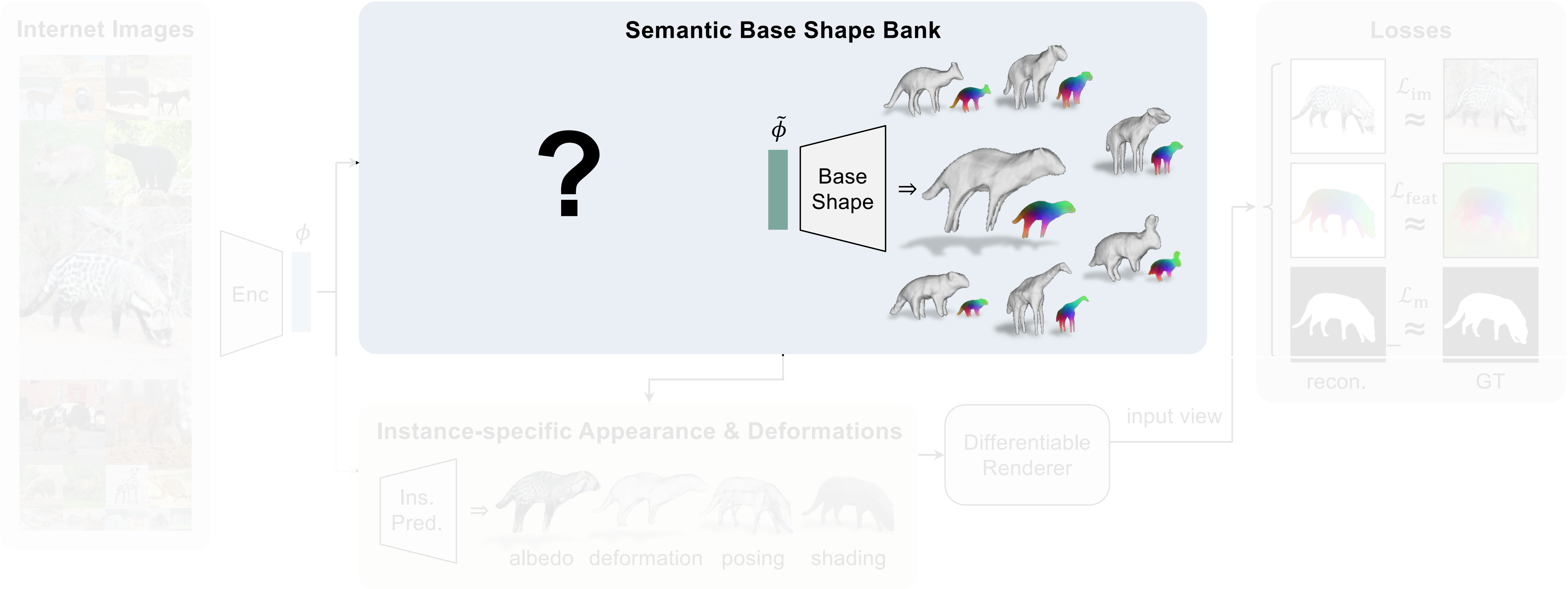


Learning the 3D Fauna of the Web. Zizhang Li, Dor Litvak, Yunzhi Zhang, Ruining Li, Tomas Jakab, Christian Rupprecht, Shangzhe Wu, Andrea Vedaldi, Jiaiun Wu. arXiv:2401.02400

Handling multiple categories

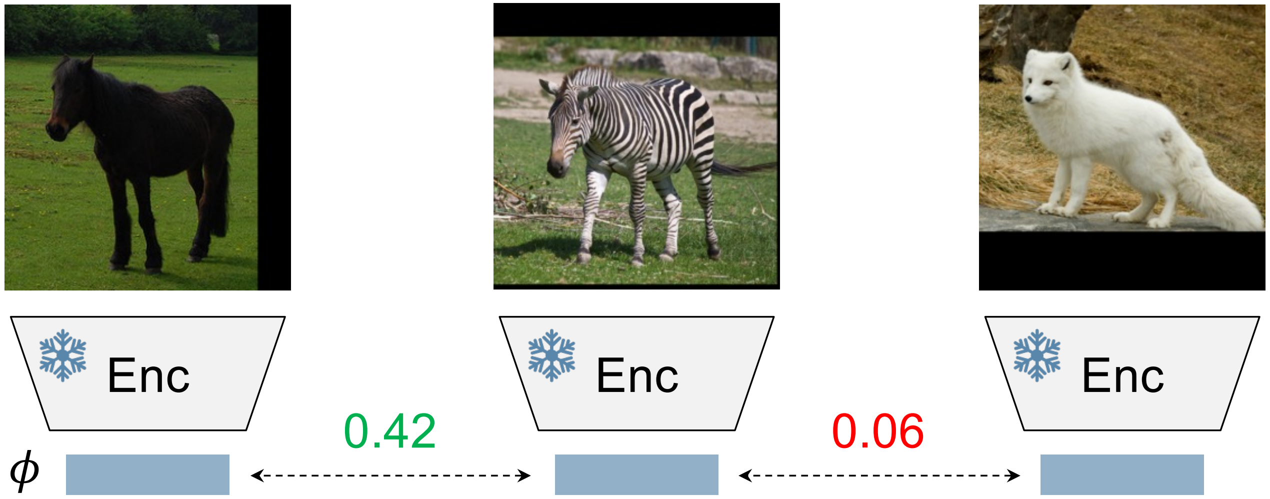


Handling multiple categories



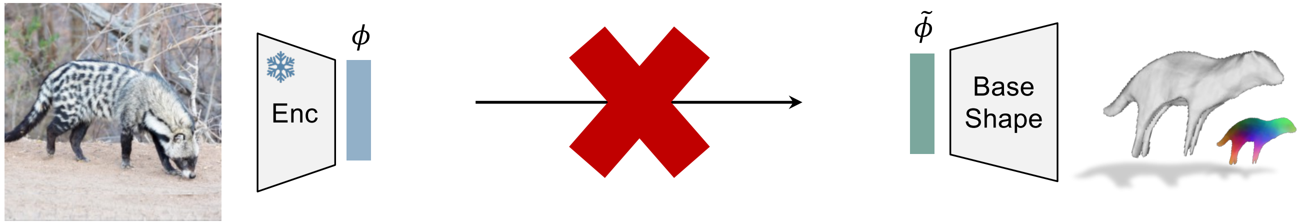
Category embedding

- Leverage a pre-trained vision encoder - DINO
  - Features serve as the soft definition of category



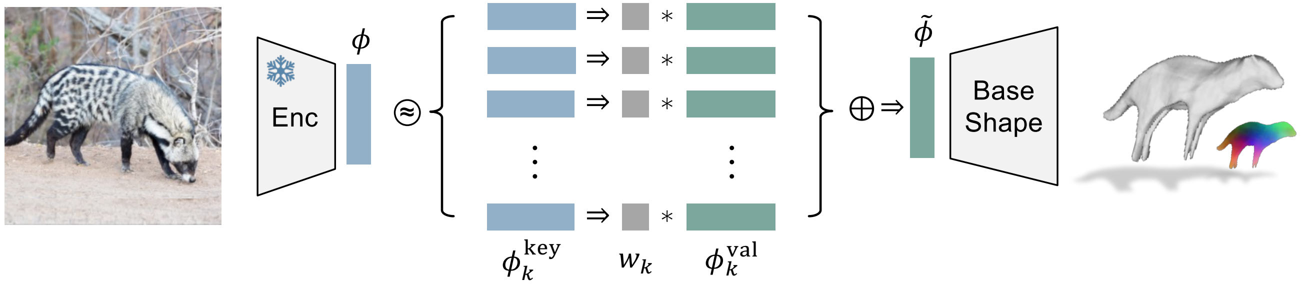
Category embedding

- Leverage a pre-trained vision encoder - DINO

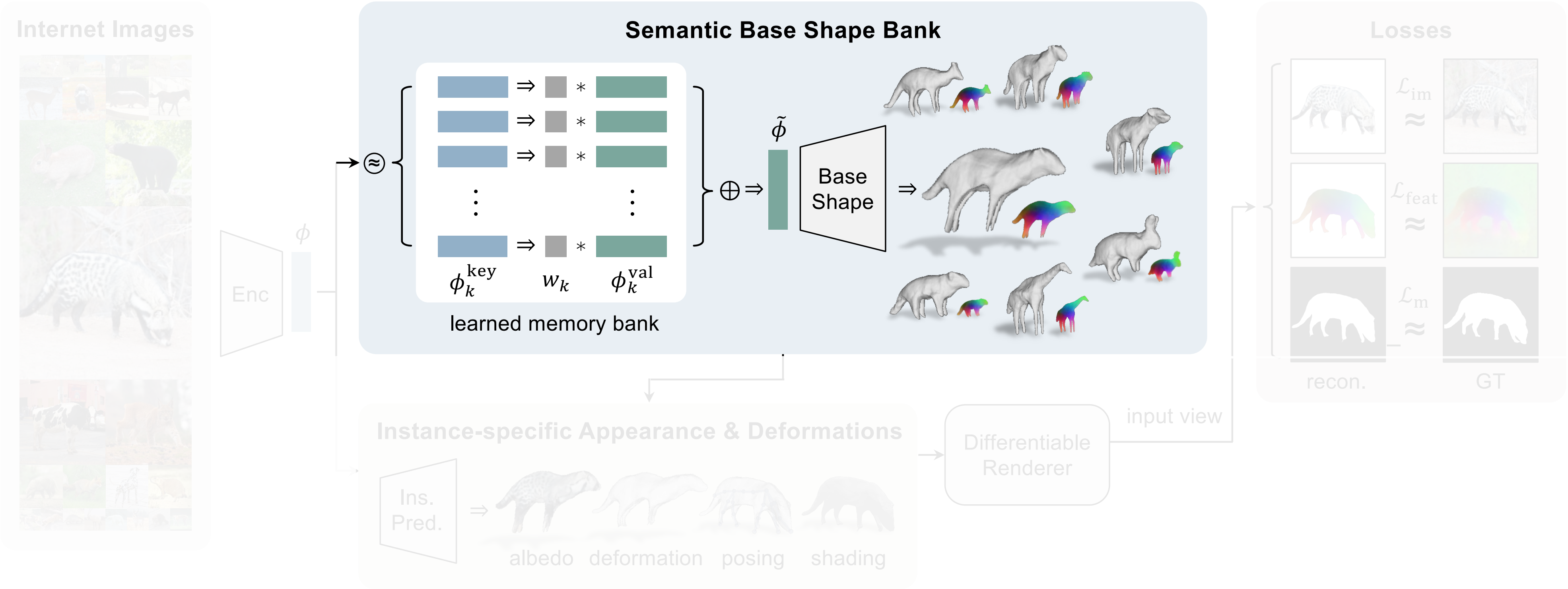


Category embedding

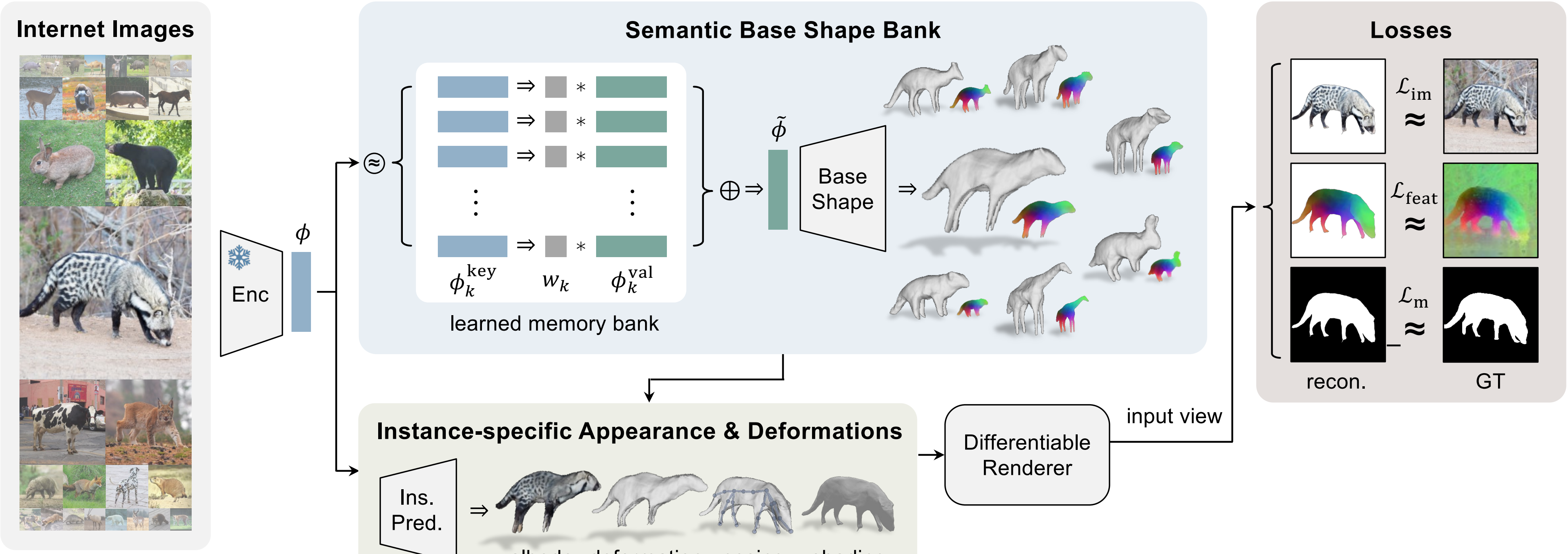
- Leverage a pre-trained vision encoder - DINO
  - A memory bank to *distills* the category information and prevents overfitting



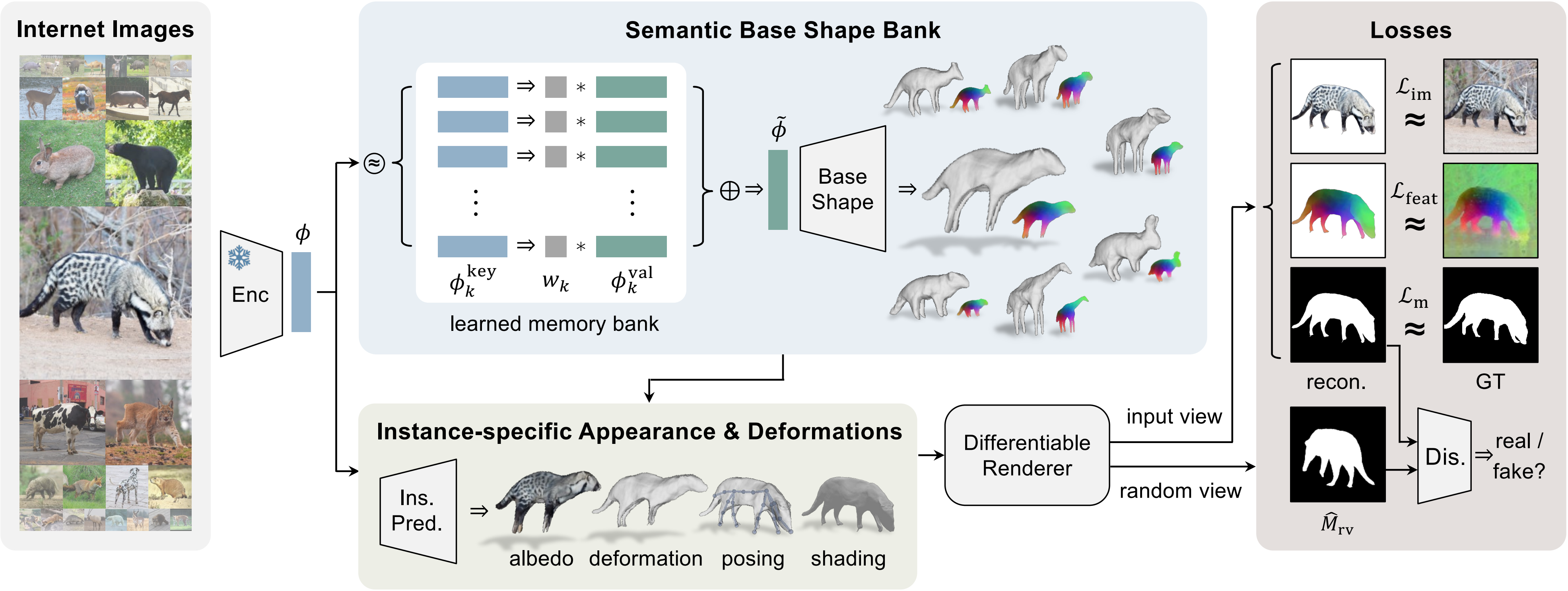
Category embedding

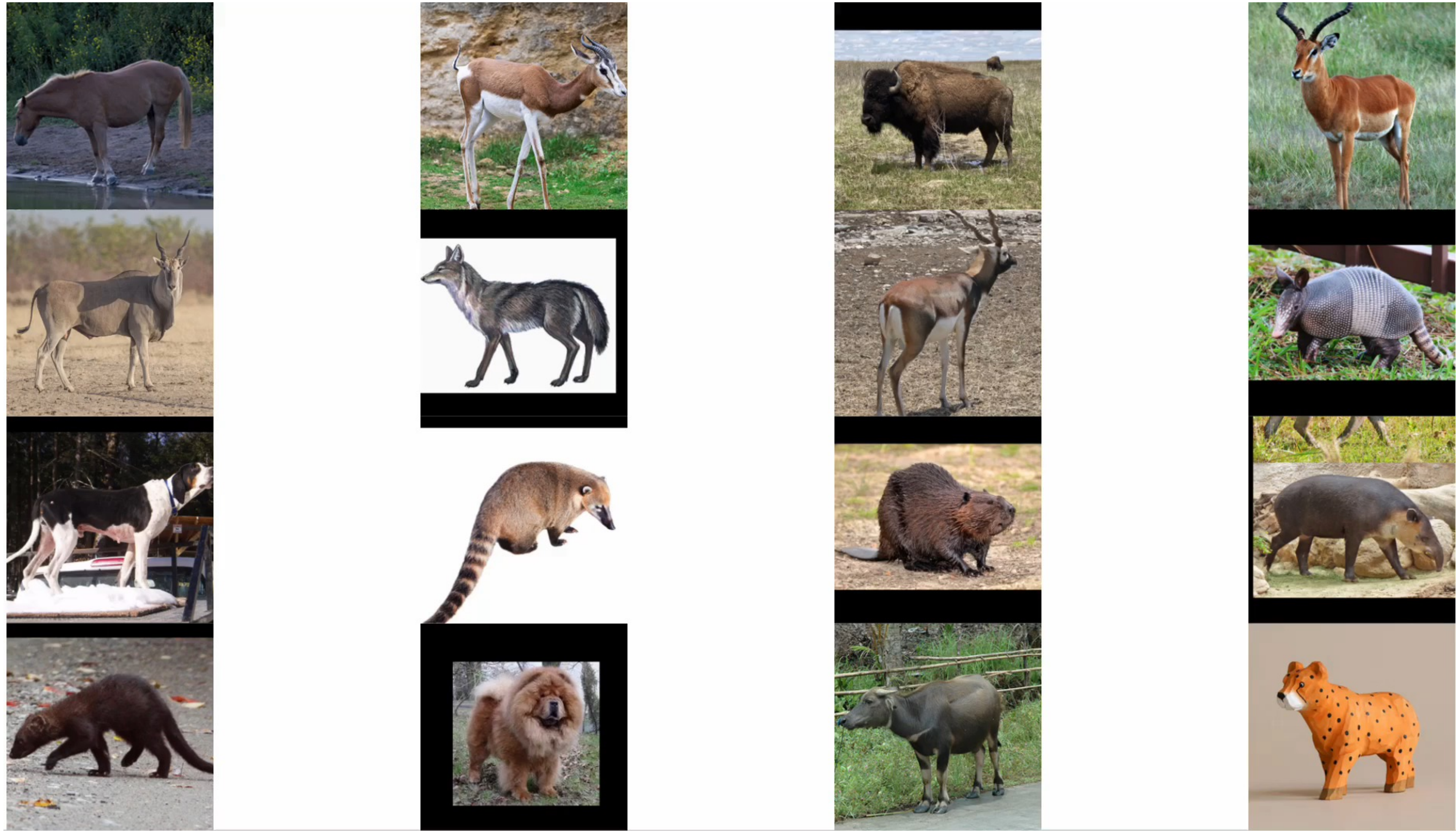


Full pipeline



Full pipeline



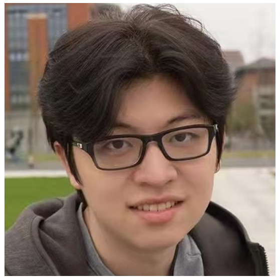


Andrea Vedaldi

Shangzhe Wu

Christian Rupprecht

Ruining Li



Zizhang Li

Dor Litvak

Yunzhi Zhang

Jiajun Wu

Learning Articulated 3D Animals from Internet Images

Tomas Jakab, University of Oxford, VGG

UNIVERSITY OF OXFORD