Learning Dynamic 3D Objects in the Wild

Elliott / Shangzhe Wu

Postdoc at Stanford SVL

Stanford University

UNIVERSITY OF OXFORD

Source: BBC Earth, https://www.youtube.com/watch?v=JWI1eCbksdE

Stable Diffusion 2.1 Demo

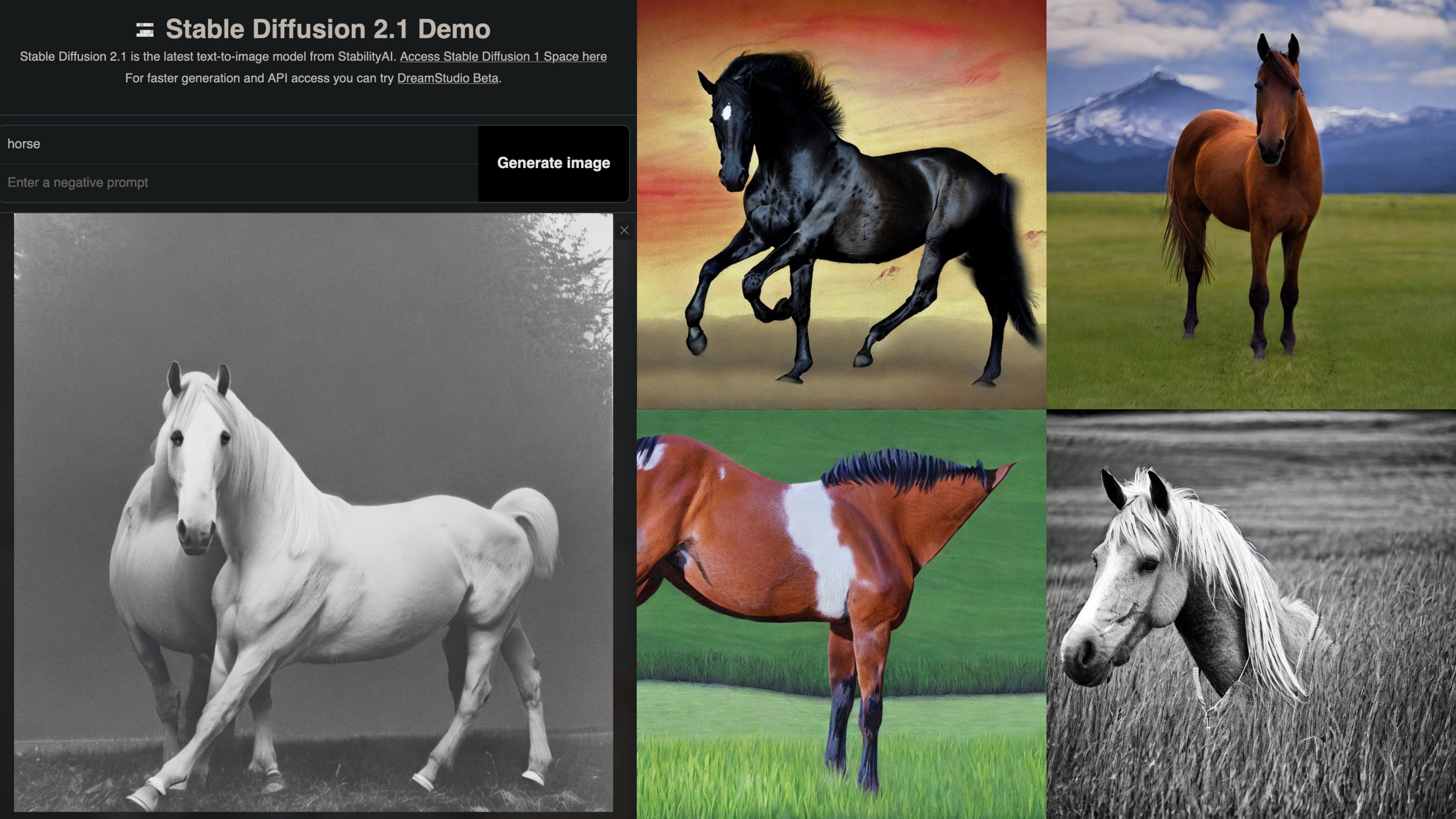
Stable Diffusion 2.1 is the latest text-to-image model from StabilityAI. Access Stable Diffusion 1 Space here

For faster generation and API access you can try DreamStudio Beta.

horse

Enter a negative prompt

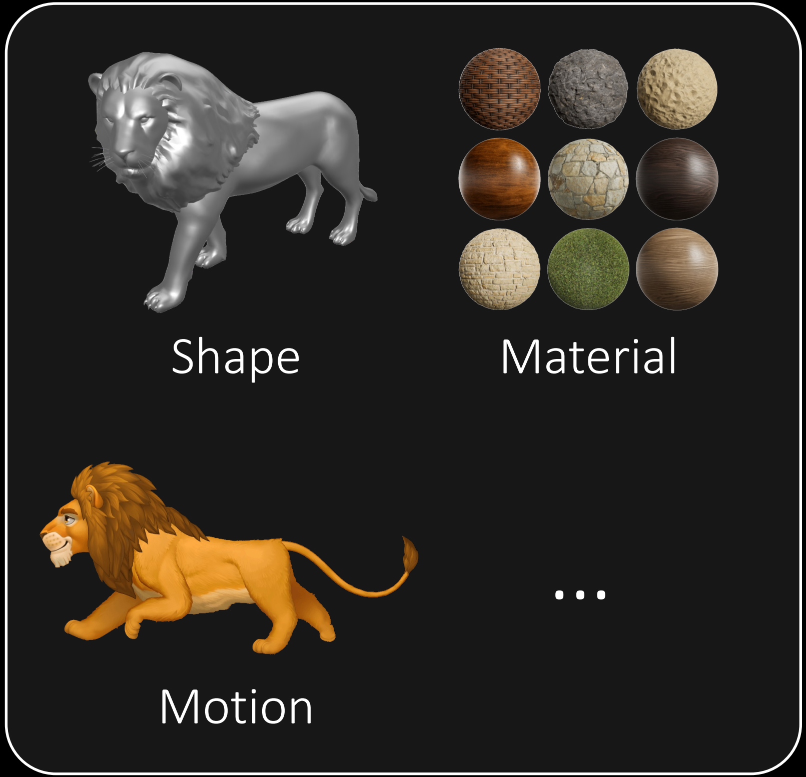
Generate image


Stable Diffusion 2.1 Demo
Stable Diffusion 2.1 is the latest text-to-image model from StabilityAI. Access Stable Diffusion 1 Space here
For faster generation and API access you can try DreamStudio Beta.
horse
Enter a negative prompt
Generate image

What is an object?

Perceiving Physical Objects beyond 2D Pixels

A “View” of an Object

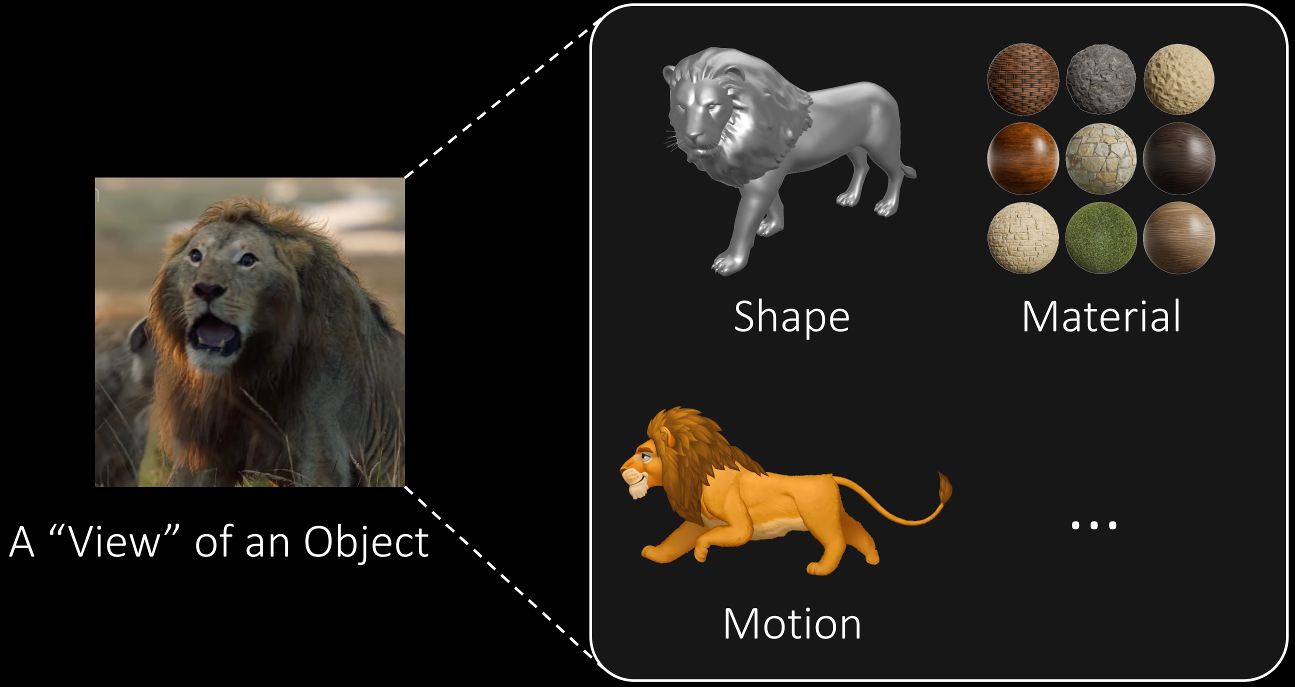


3D Object Priors

Geometric Annotations by Humans

COCO
Common Objects in Context

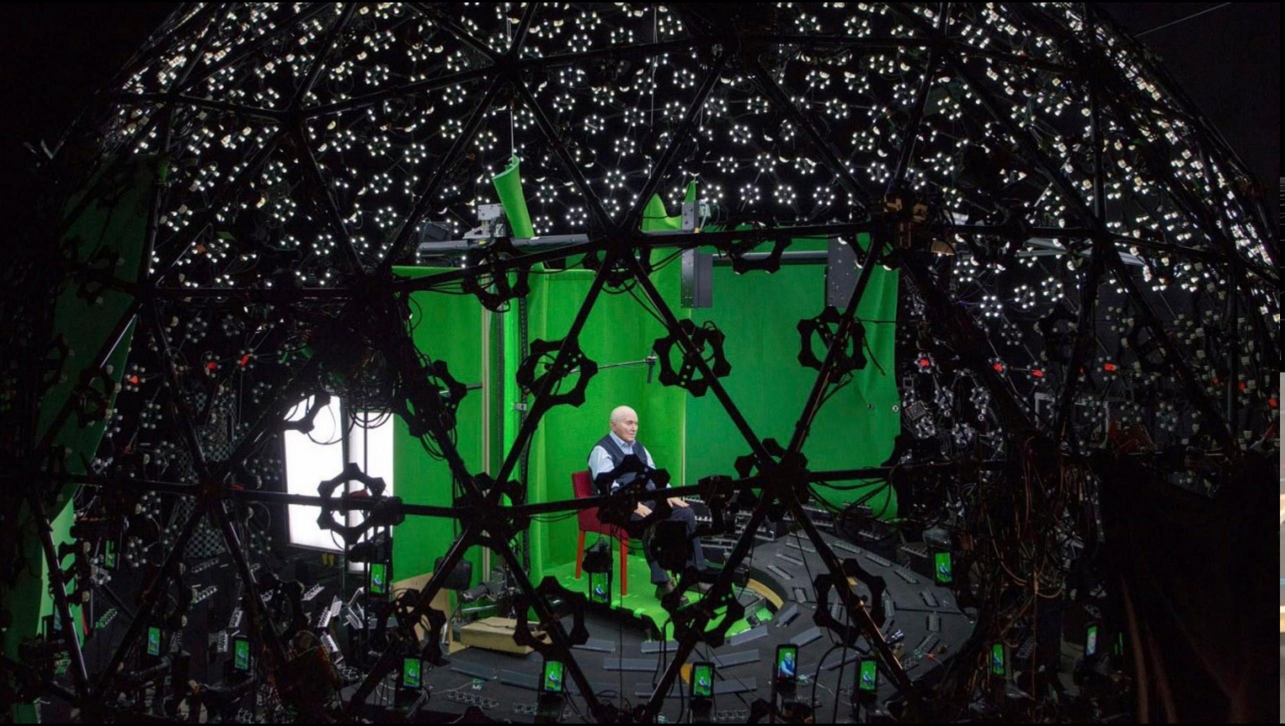
Annotation beyond 2D is hard!

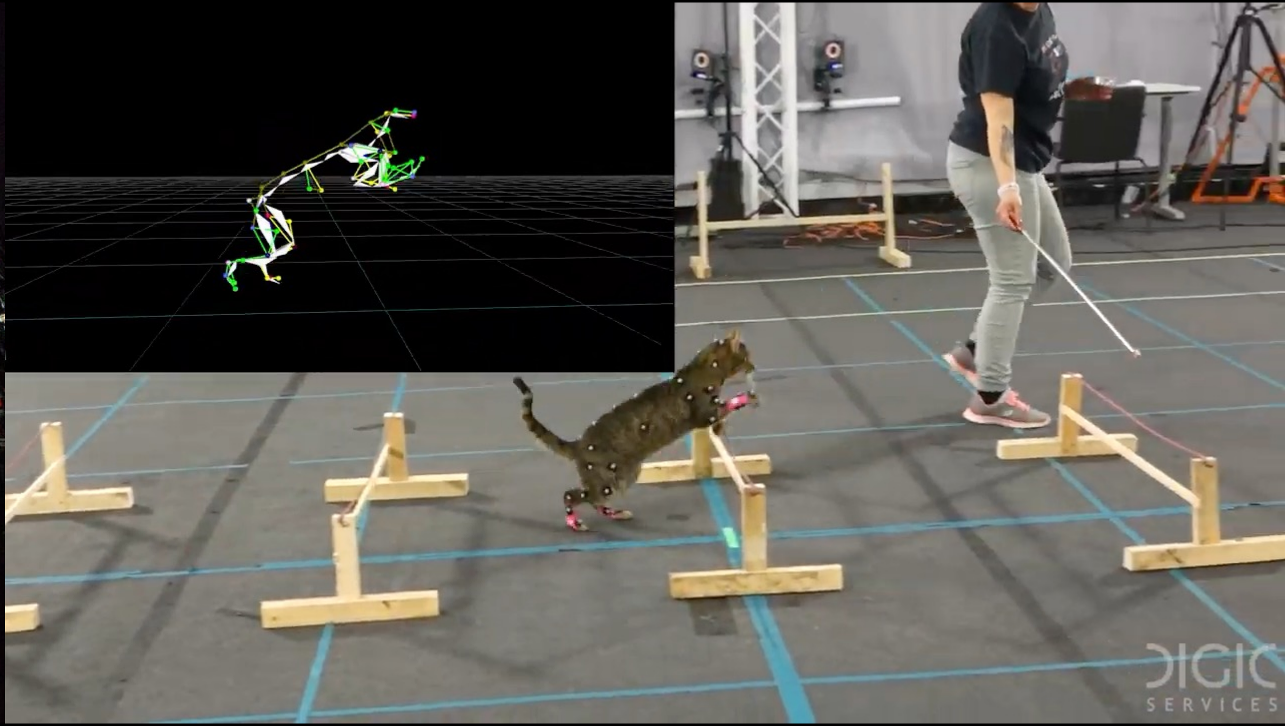


Physically-grounded 3D Representations

- 3D surfaces, normals?
- Materials (BRDFs)?
- Environment lighting?
- Physics: force, torque, mass, friction, velocity, acceleration…?

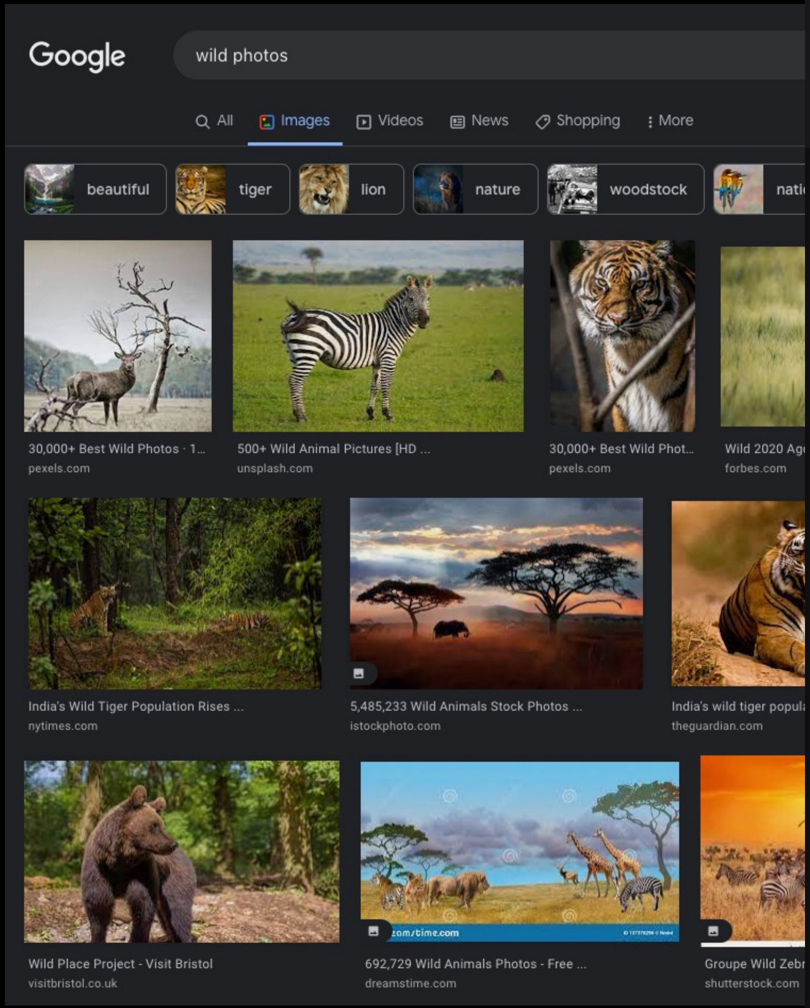
Special Capturing Devices

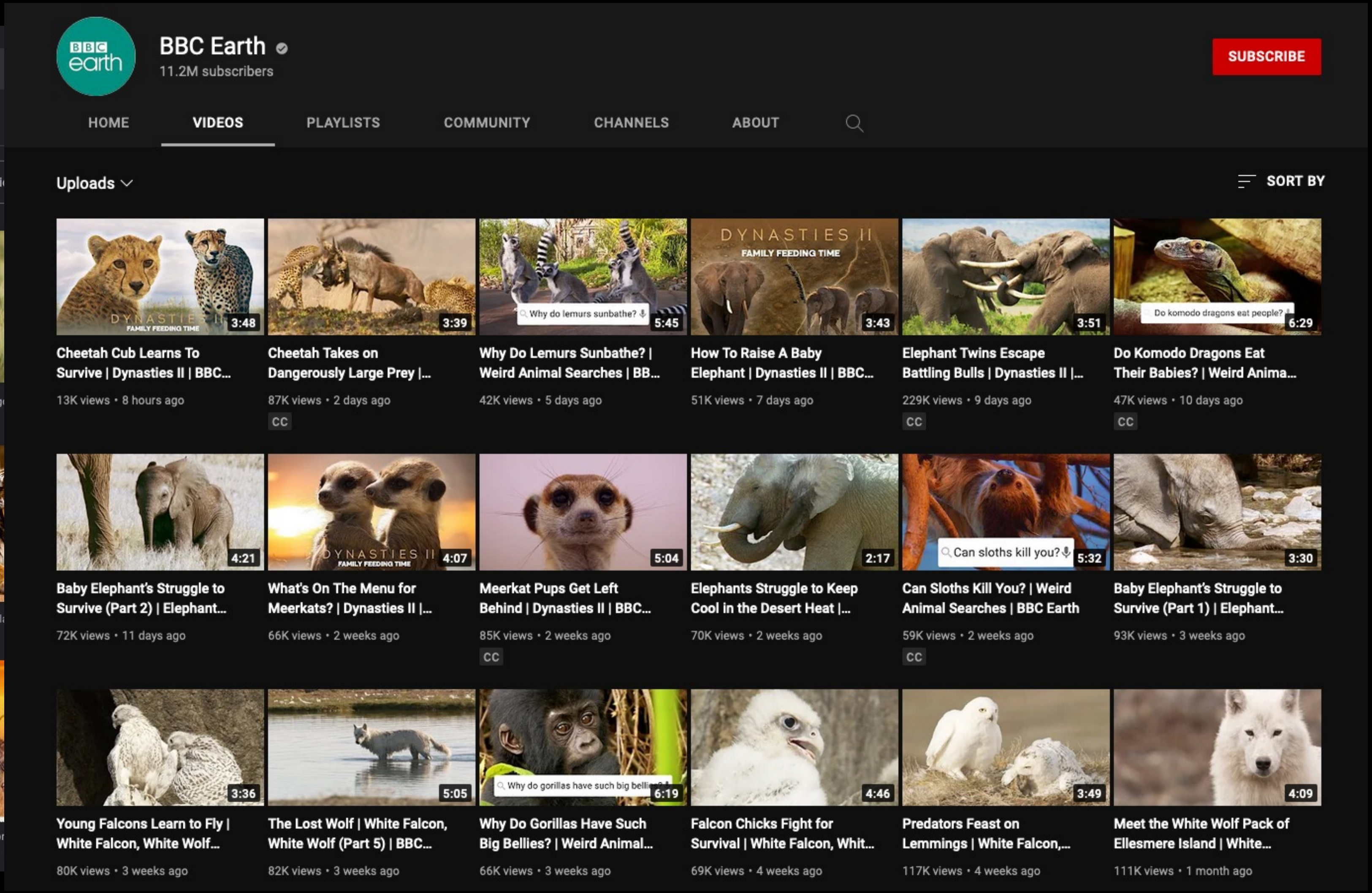




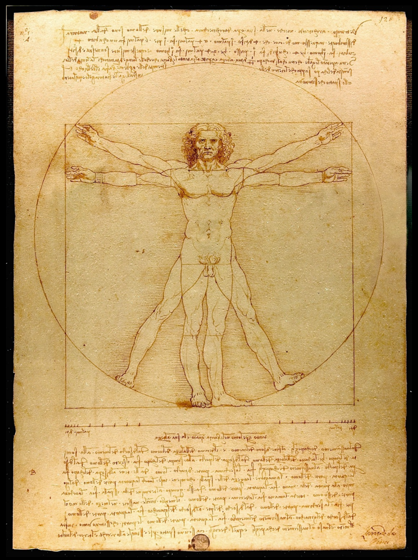
Hard to scale up to all kinds of objects

Can we simply learn from “the wild”?





Luckily, we know how the world works (at least kind of…)



- It's a physical 3D world
- Lots of symmetries / regularities
- We can simulate the image formation process
- …

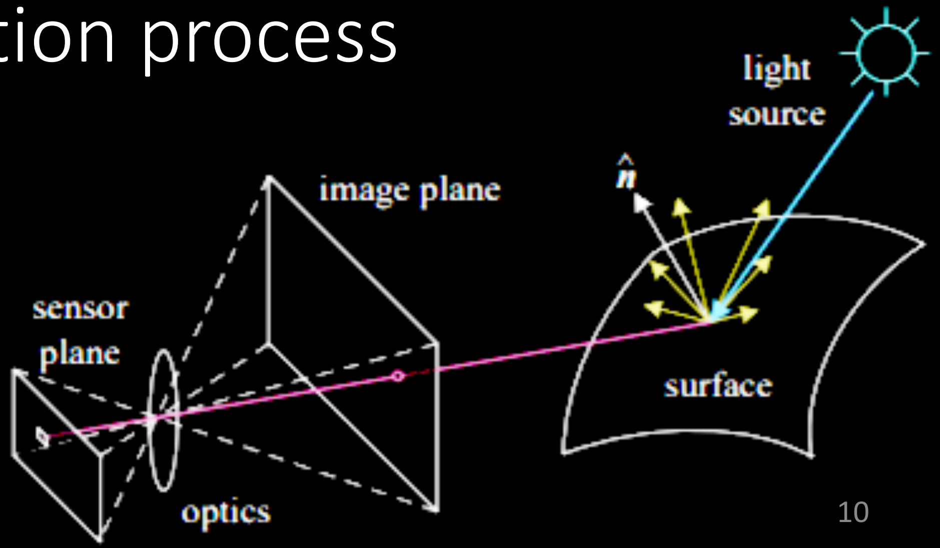
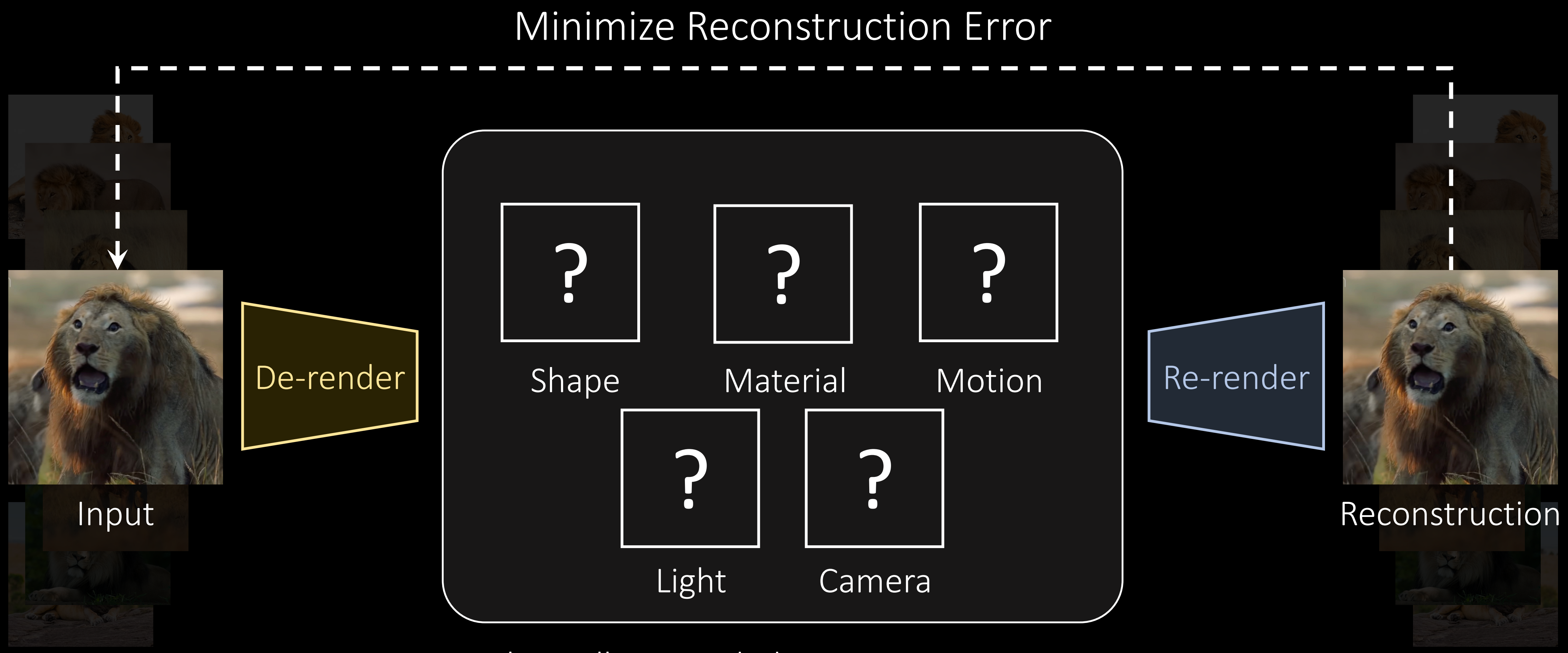
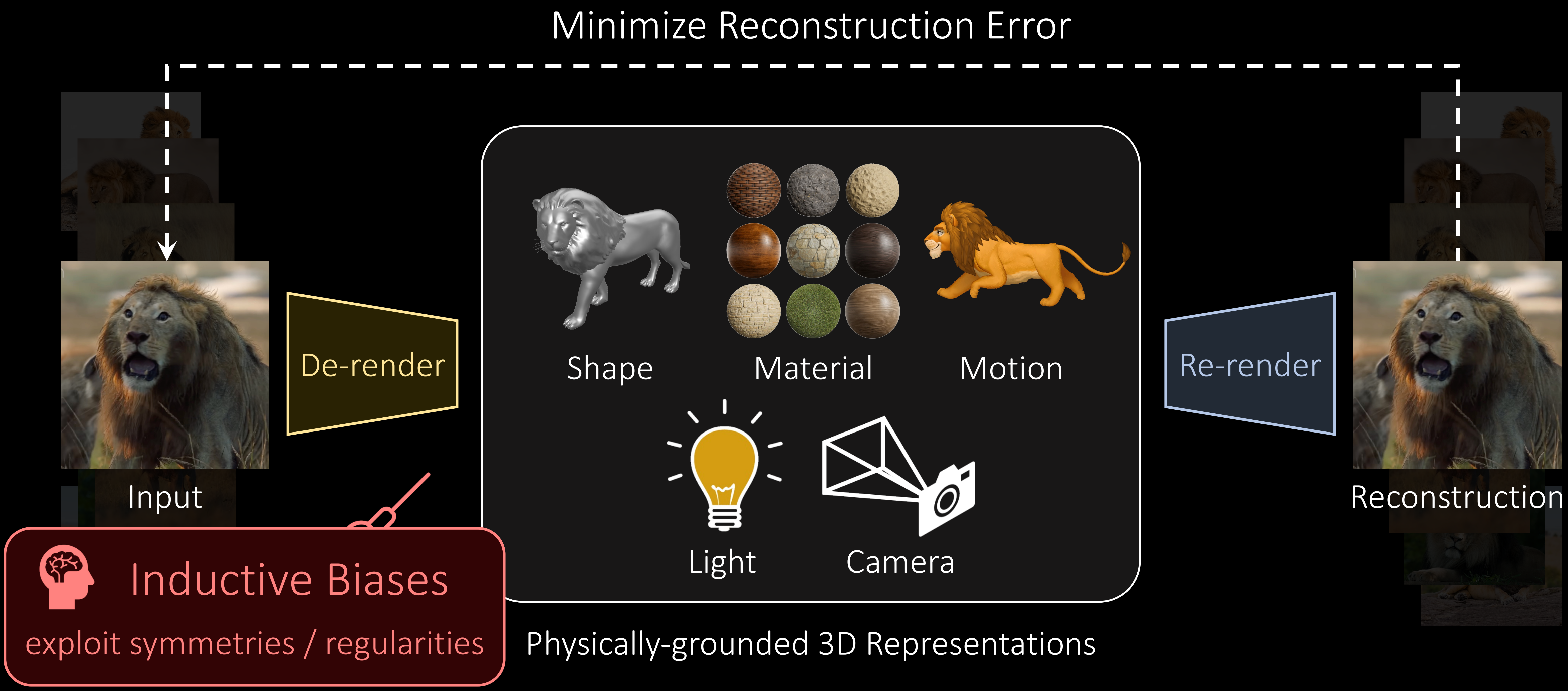


Photo-Geometric Autoencoding



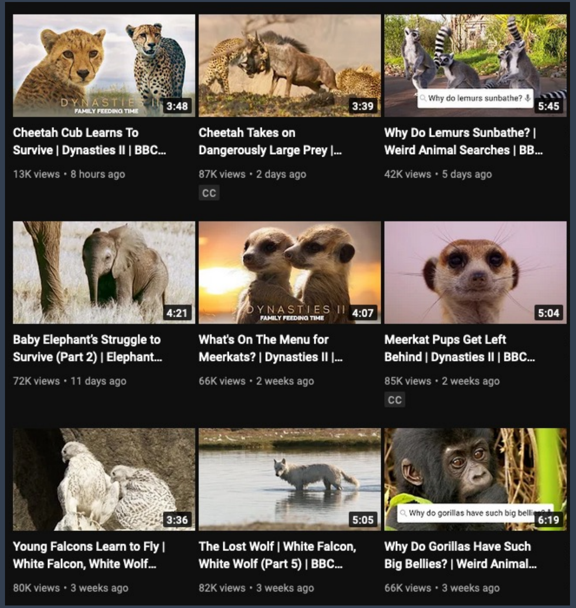
Physically-grounded 3D Representations

Photo-Geometric Autoencoding



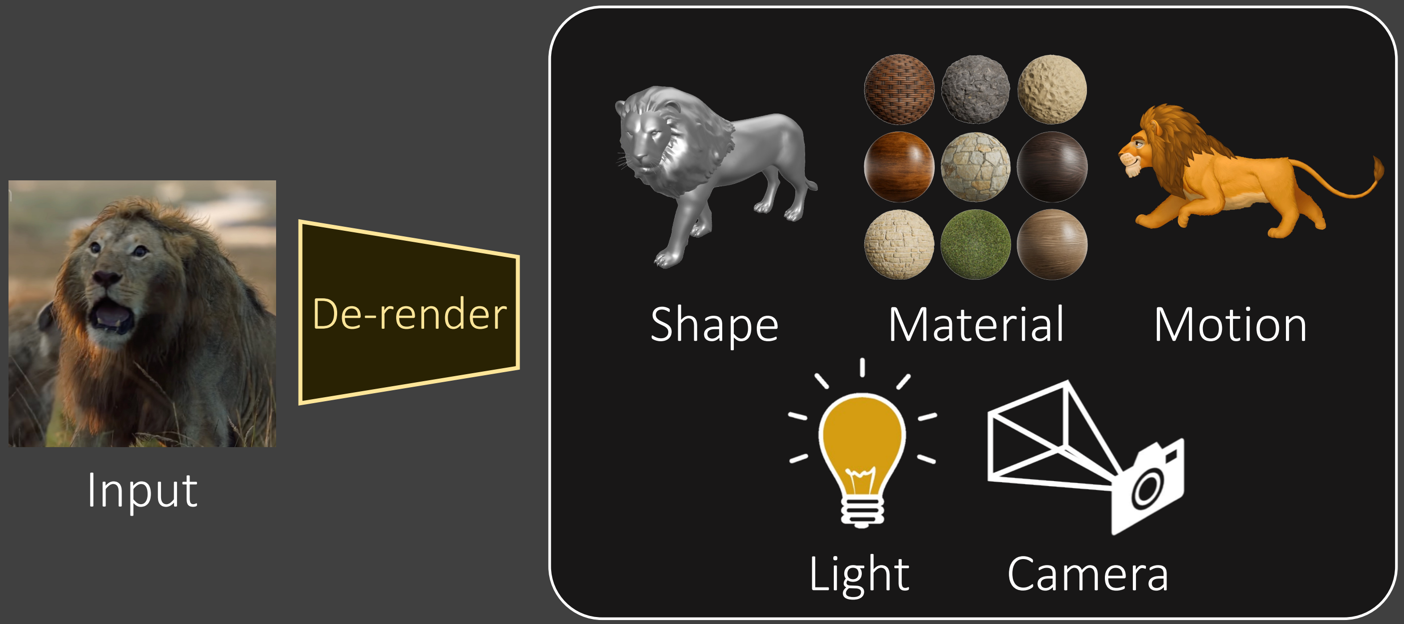
Learning Physical 3D Objects in the Wild

Training



"In-the-Wild" Data

Inference – Single Image De-rendering



Physically-grounded 3D Representations

Physics offers a path for learning compact, generalizable object representations.

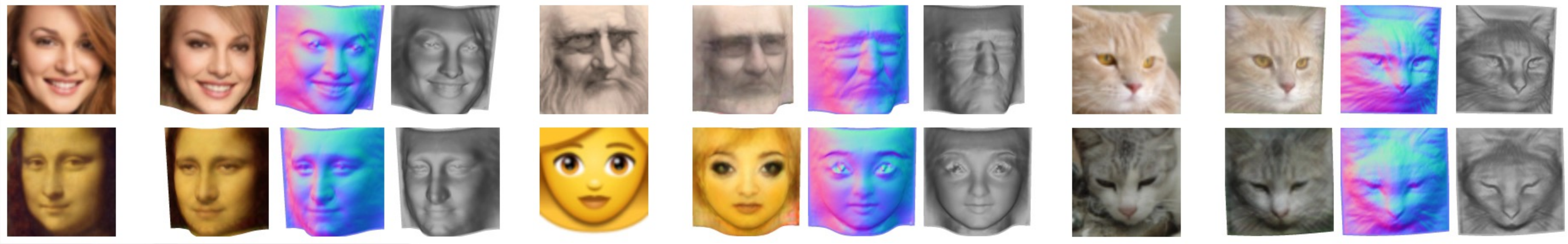
Unsupervised 3D Learning in the Wild

- 3D annotations are expensive and often infeasible at scale.
- Towards first principles in vision:
  - What are the minimal assumptions for 3D perception?
- Learning through inverse rendering gives rise to:
  - Physical interpretability and verifiability
  - Better generalization
  - Controllable generation

Unsupervised Learning of Probably Symmetric Deformable 3D Objects from Images in the Wild

CVPR 2020

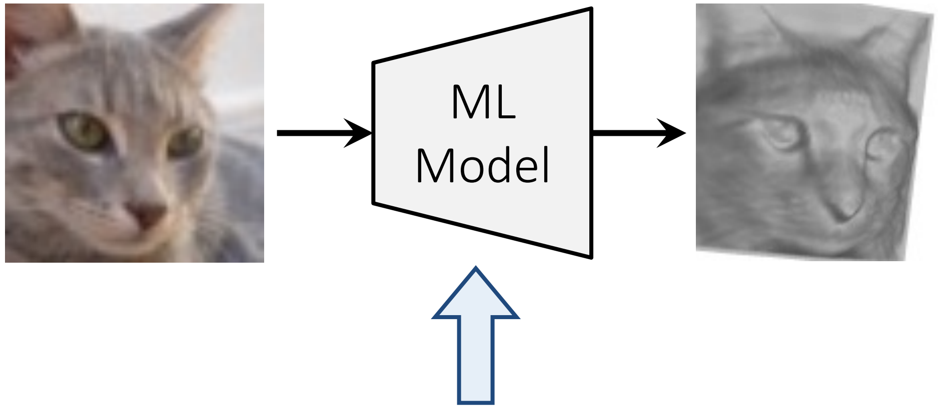
Shangzhe Wu   Christian Rupprecht   Andrea Vedaldi



VGG University of Oxford

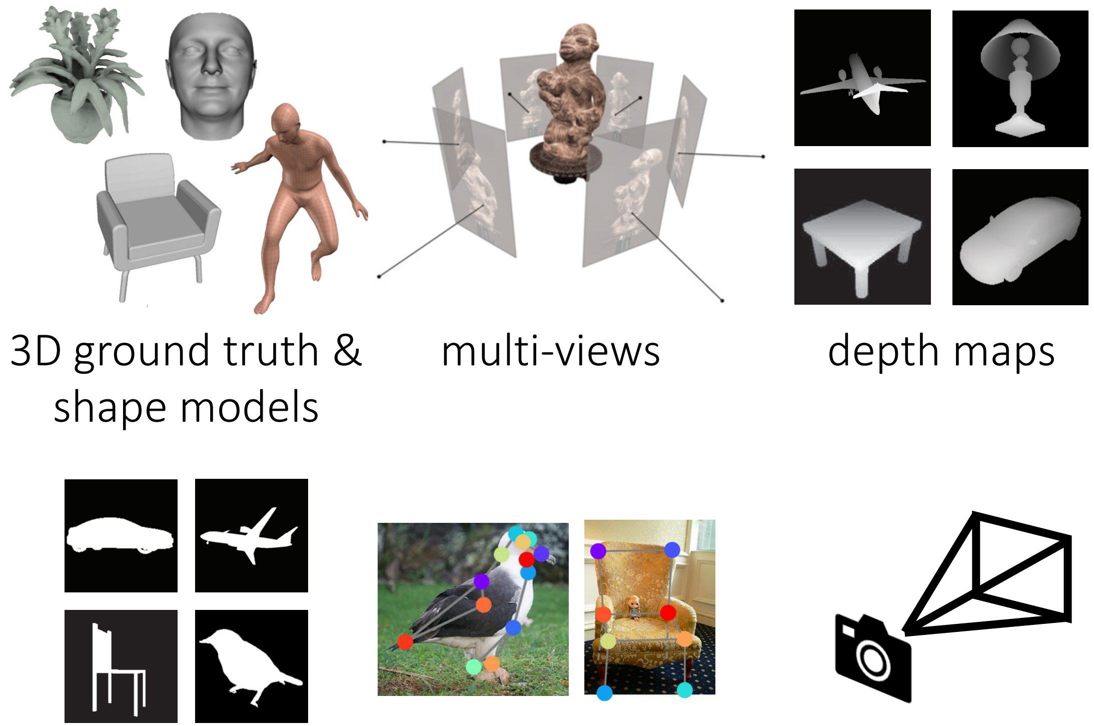
UNIVERSITY OF OXFORD

Learning-based Single-view 3D Reconstruction



3D priors learned during training

Supervision



3D ground truth & shape models

multi-views

depth maps

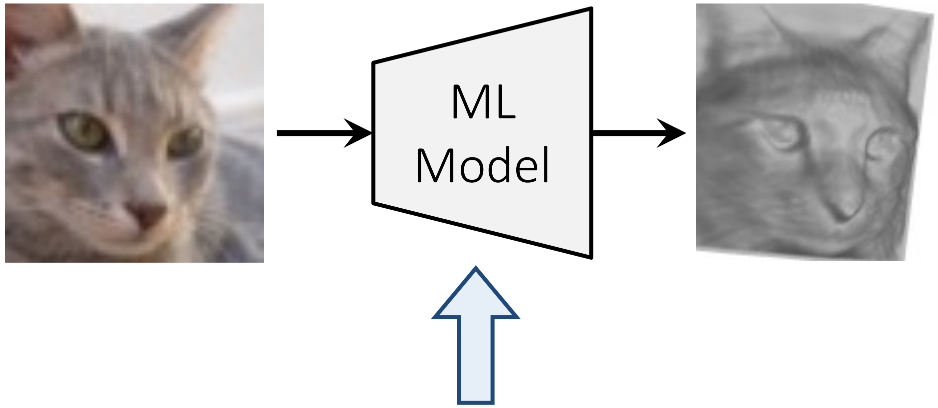
silhouettes

keypoints

camera viewpoint

3D annotations are expensive!

Unsupervised Single-view 3D Reconstruction

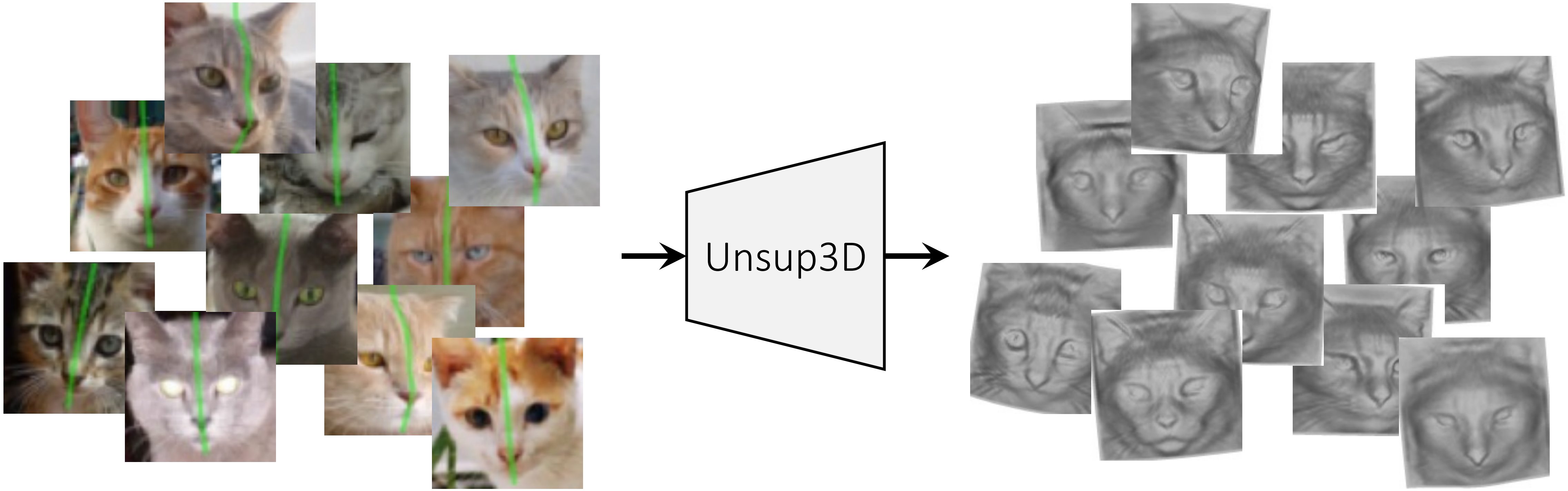


3D priors learned during training

Unsupervised Learning of Symmetric 3D Objects

Training Data

Output



single-view images of a category

NO other supervision!

single image 3D reconstruction

Photo-Geometric Autoencoding

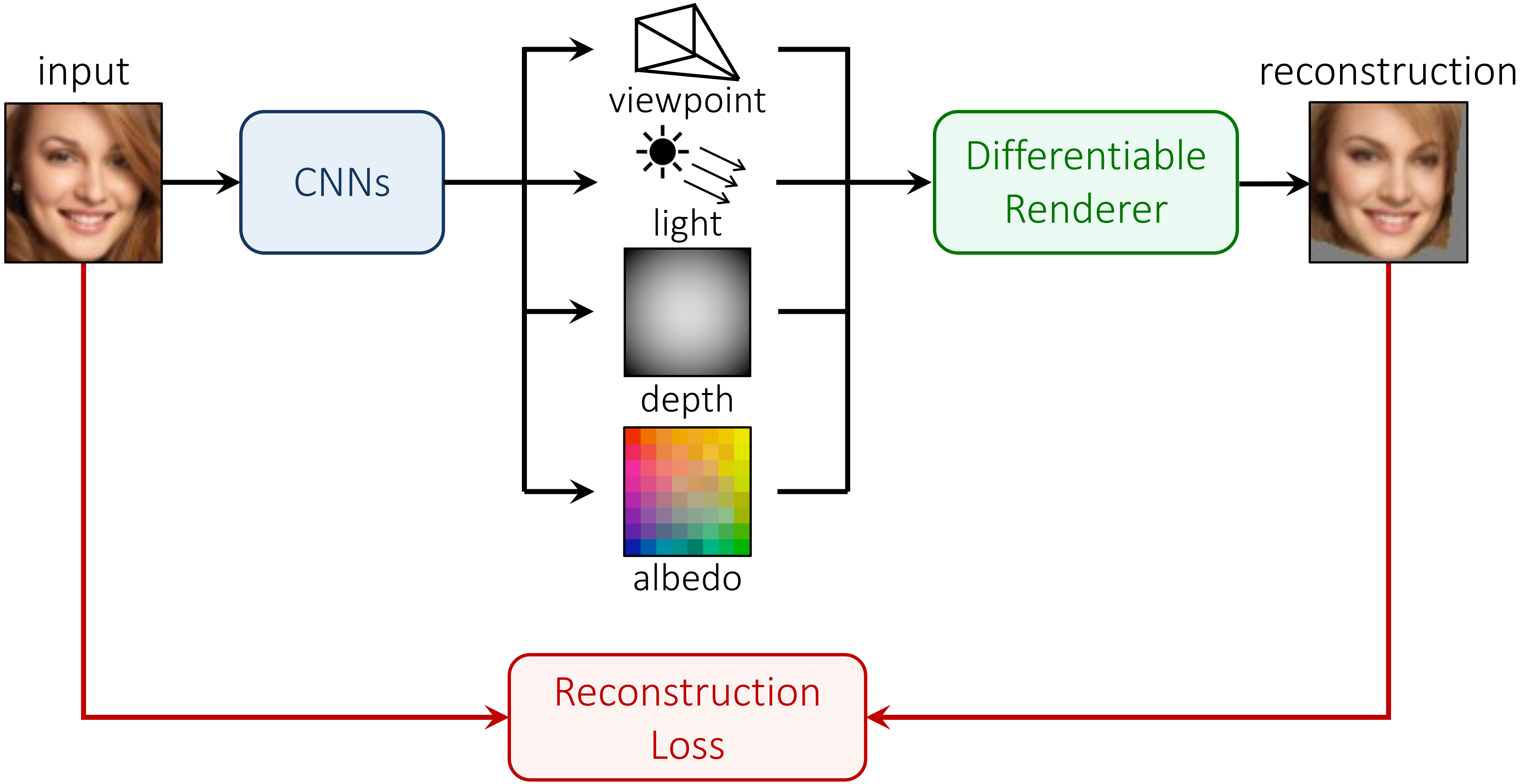


Photo-Geometric Autoencoding

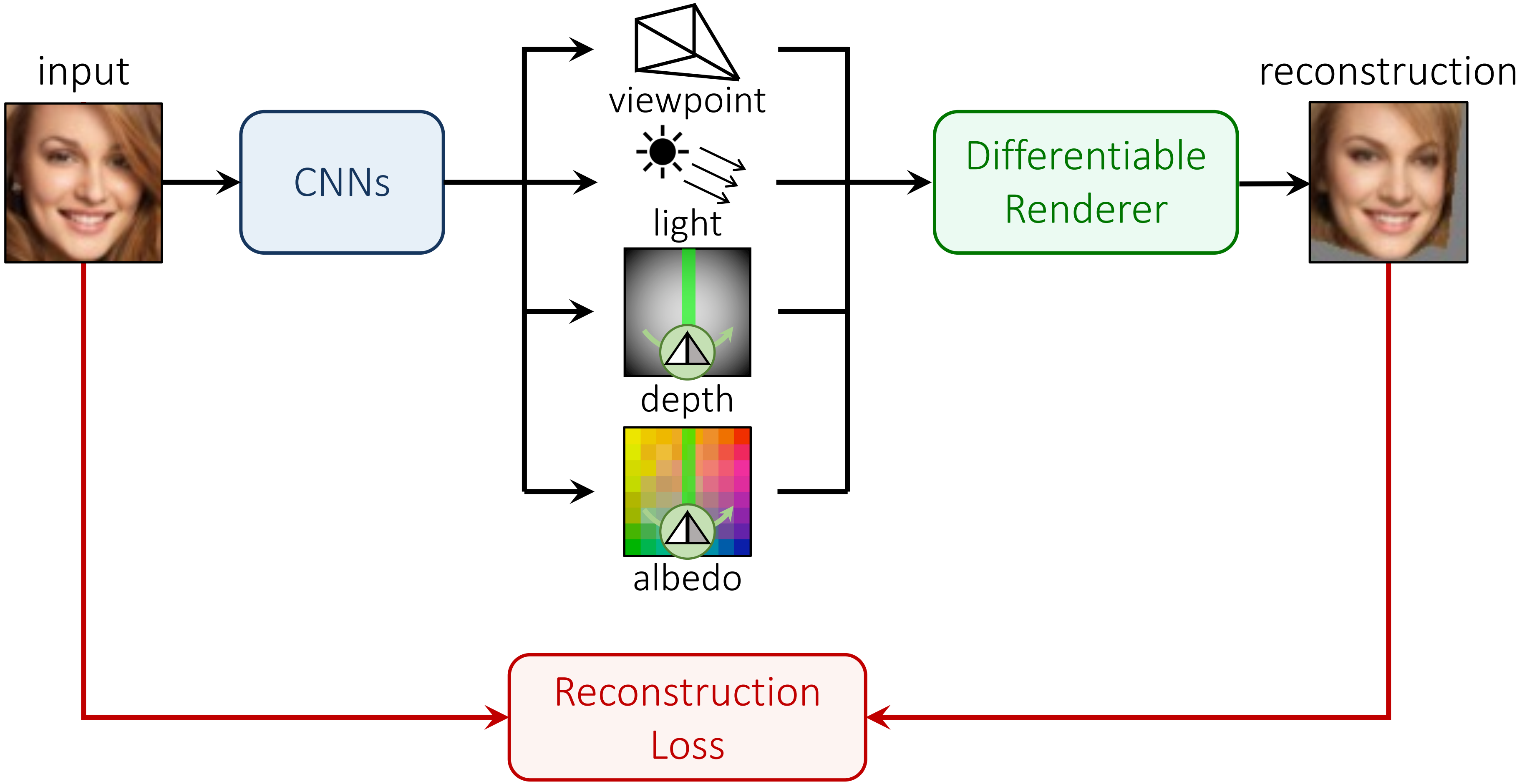
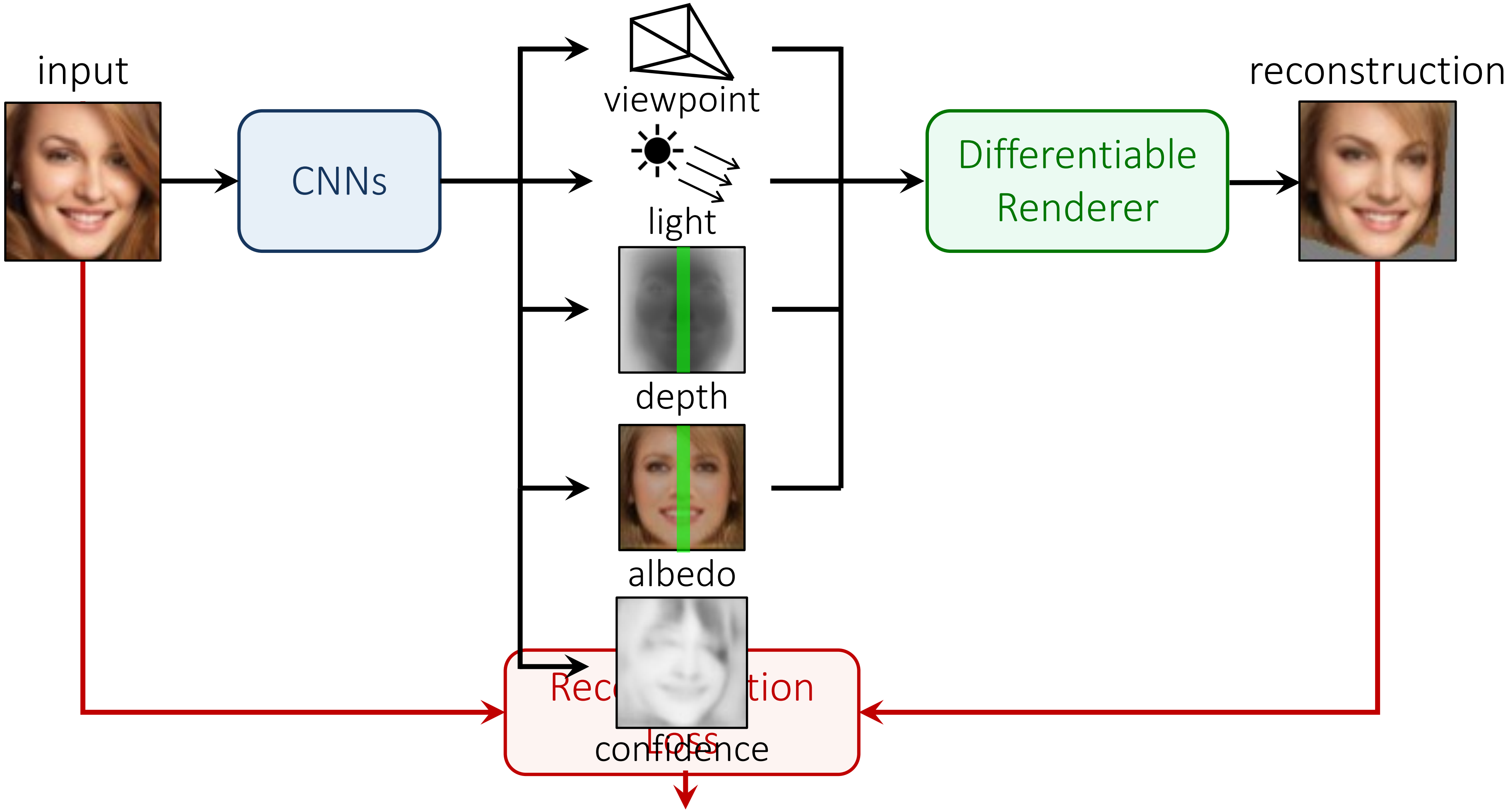
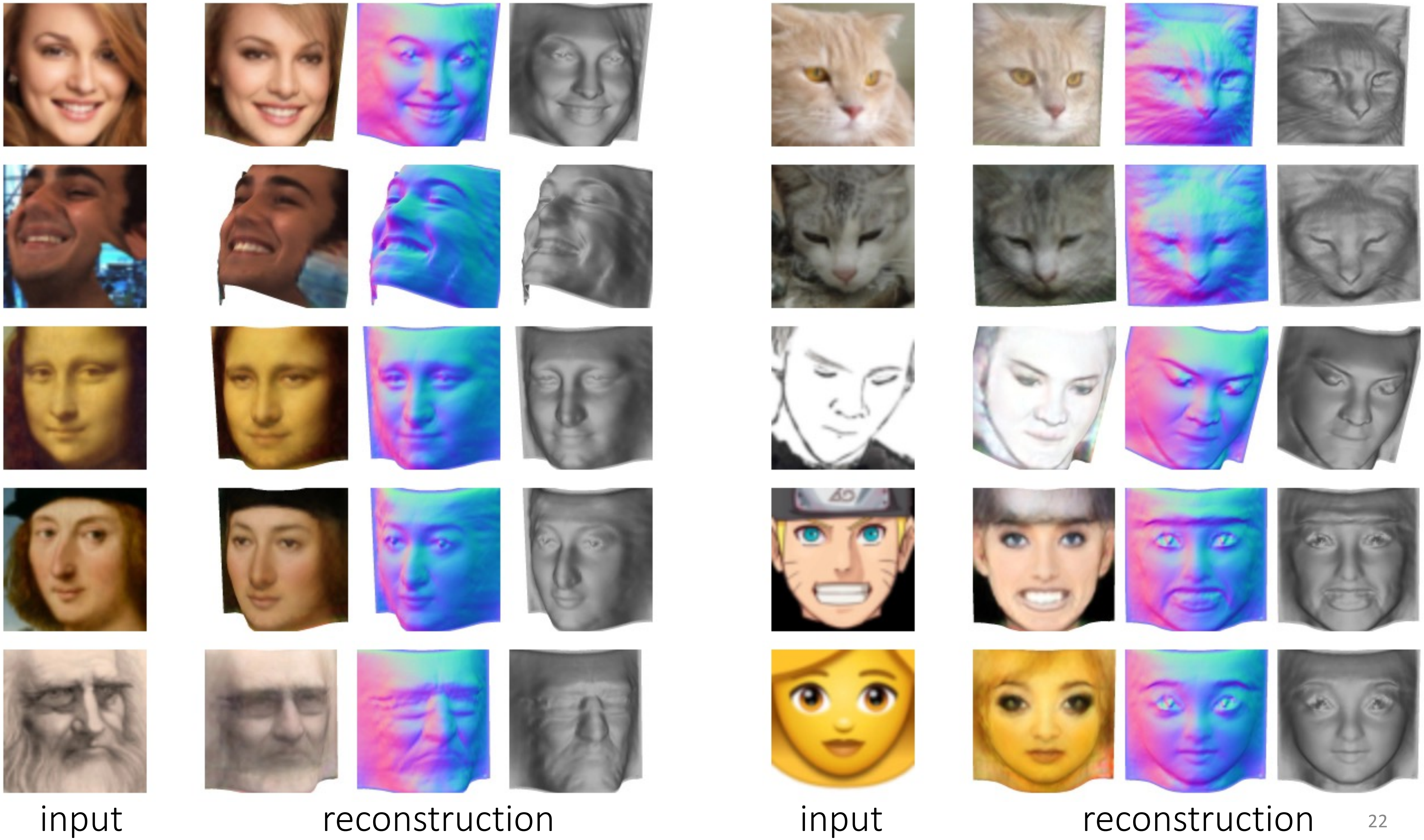
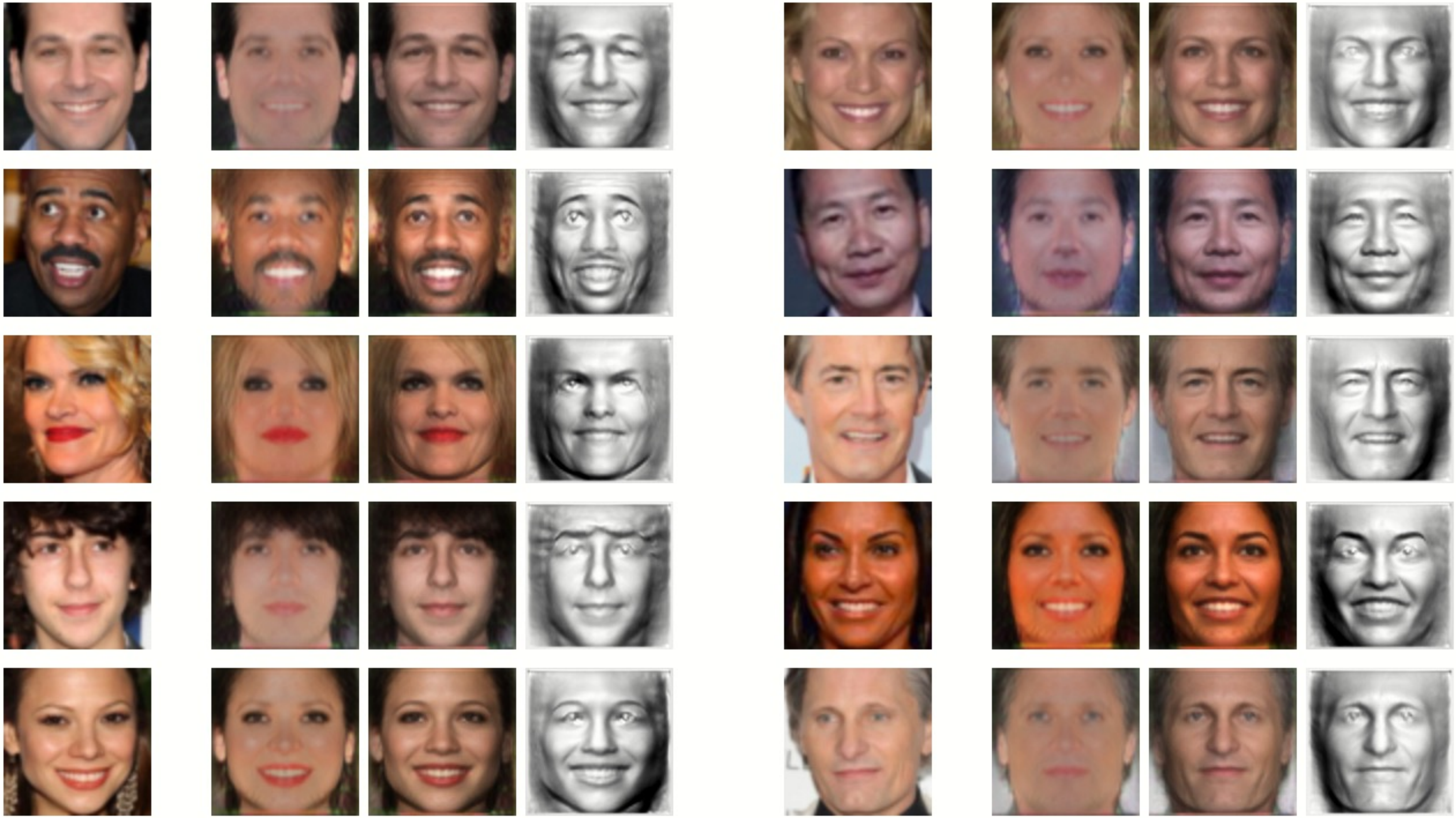


Photo-Geometric Autoencoding with Symmetry




input
reconstruction
input
reconstruction



input decompose & relight input decompose & relight

MagicPony: Learning Articulated 3D Animals in the Wild

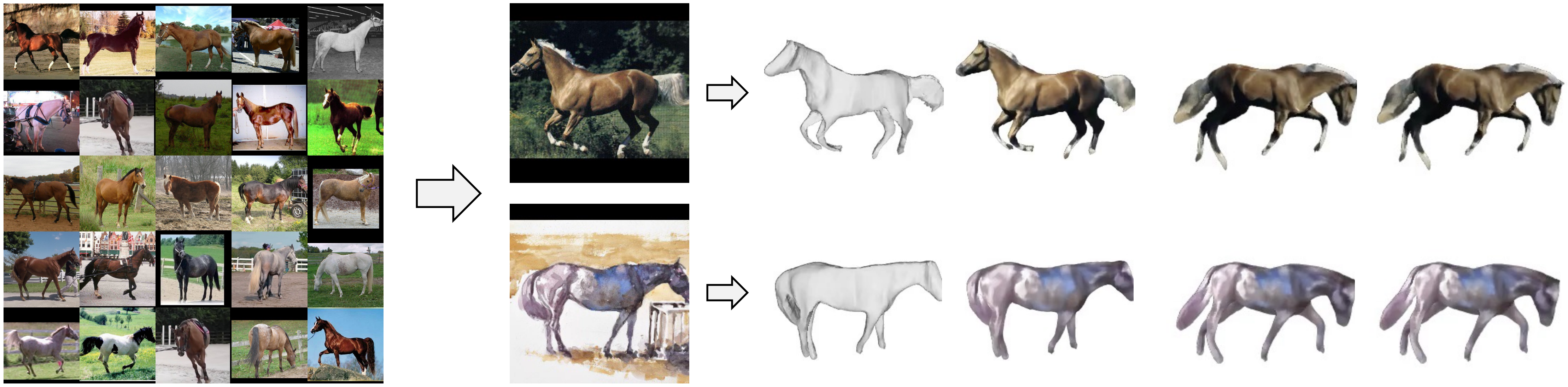
Shangzhe Wu*   Ruining Li*   Tomas Jakab*   Christian Rupprecht   Andrea Vedaldi

Visual Geometry Group, University of Oxford

(* Equal Contribution)

CVPR 2023

Training **Single-Image Inference**



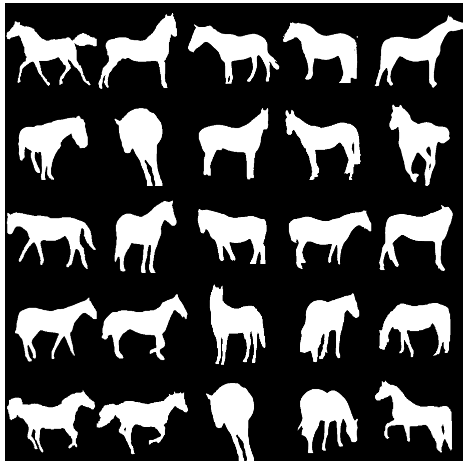
Single-view Images   Test Image   Articulated 3D Shape   Animation

Training Data

Single-view Images

Off-the-shelf PointRend [1]

Off-the-shelf DINO-ViT [2]



Instance Masks

Self-supervised Image Features

No keypoint or viewpoint supervision, nor template shapes

[1] PointRend: Image Segmentation as Rendering. Kirillov et. al. CVPR 2020. [2] Emerging Properties in Self-supervised Vision Transformers. Caron et. al. ICCV 2021.

Correspondences from Self-supervised DINO Features

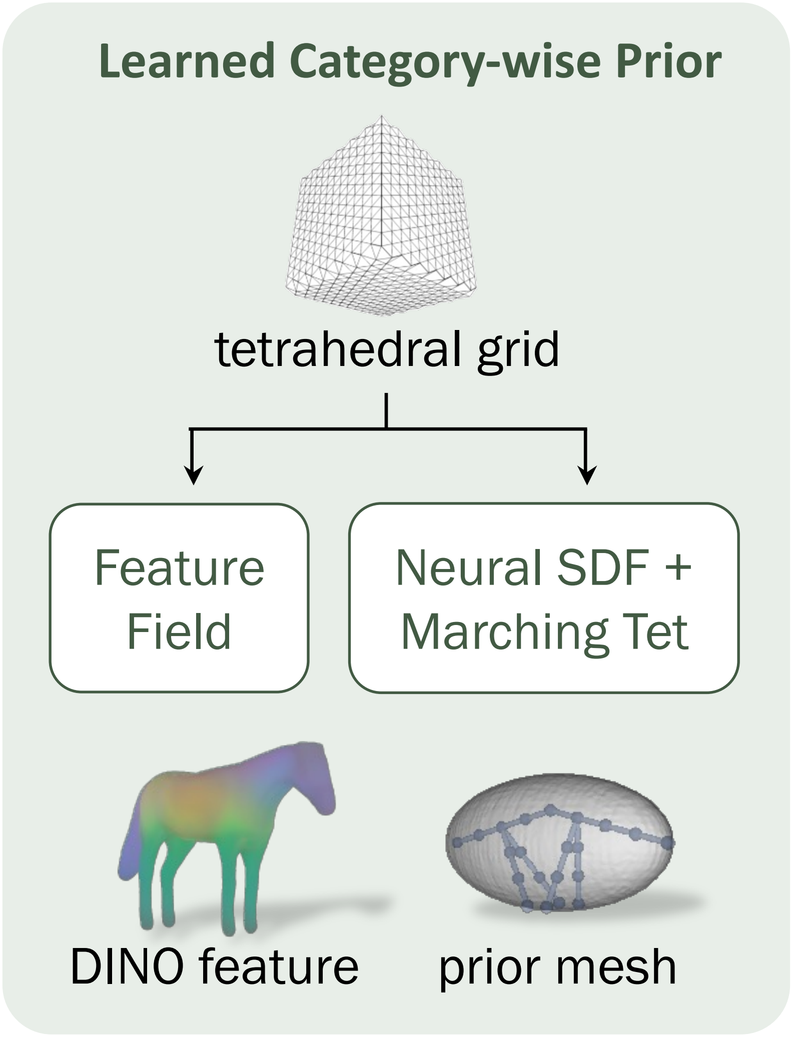
Self-supervised Image Features

Correspondences from Self-supervised DINO Features

Learned Category-wise Prior

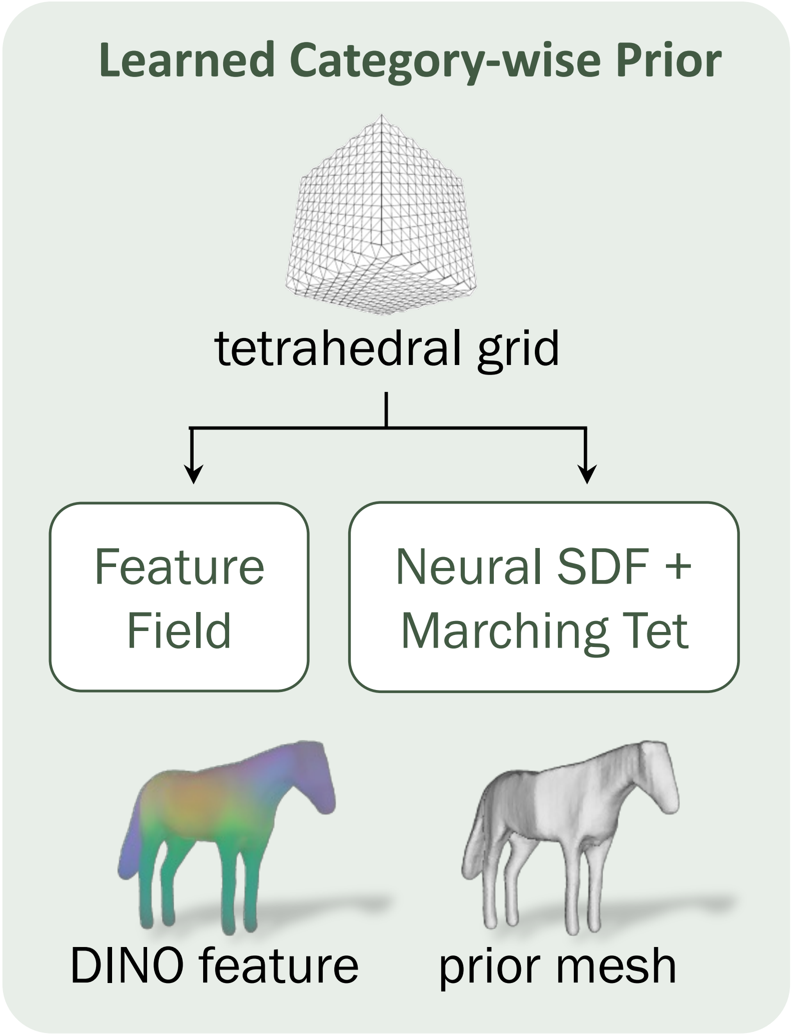
learned canonical DINO feature

Implicit-Explicit 3D Representation



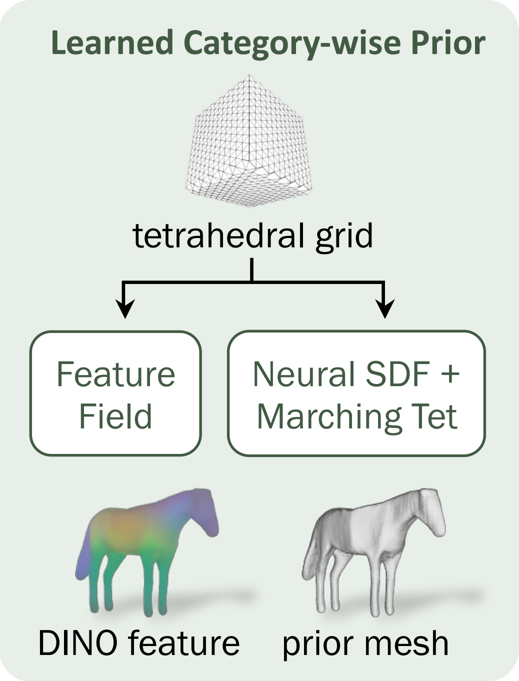
[1] Deep Marching Tetrahedra: a Hybrid Representation for High-Resolution 3D Shape Synthesis. Shen et. al. NeurIPS 2021.

Implicit-Explicit 3D Representation



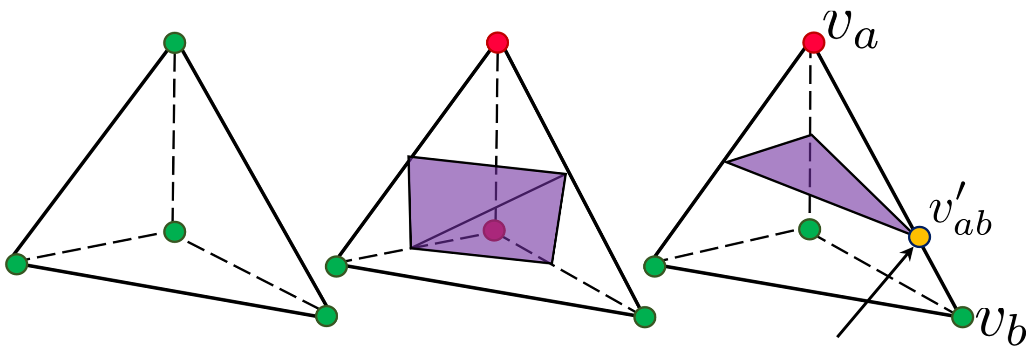
[1] Deep Marching Tetrahedra: a Hybrid Representation for High-Resolution 3D Shape Synthesis. Shen et. al. NeurIPS 2021.

Implicit-Explicit 3D Representation



Deep Marching Tetrahedra (DMTet)

Triangular meshes from Signed Distance Function (SDF) $s(\cdot)$



- positive SDF
- negative SDF

$$v'_{ab} = \frac{v_a \cdot s(v_b) - v_b \cdot s(v_a)}{s(v_b) - s(v_a)}$$

SDF
✓ Flexible topology
✓ Smooth gradients

\+

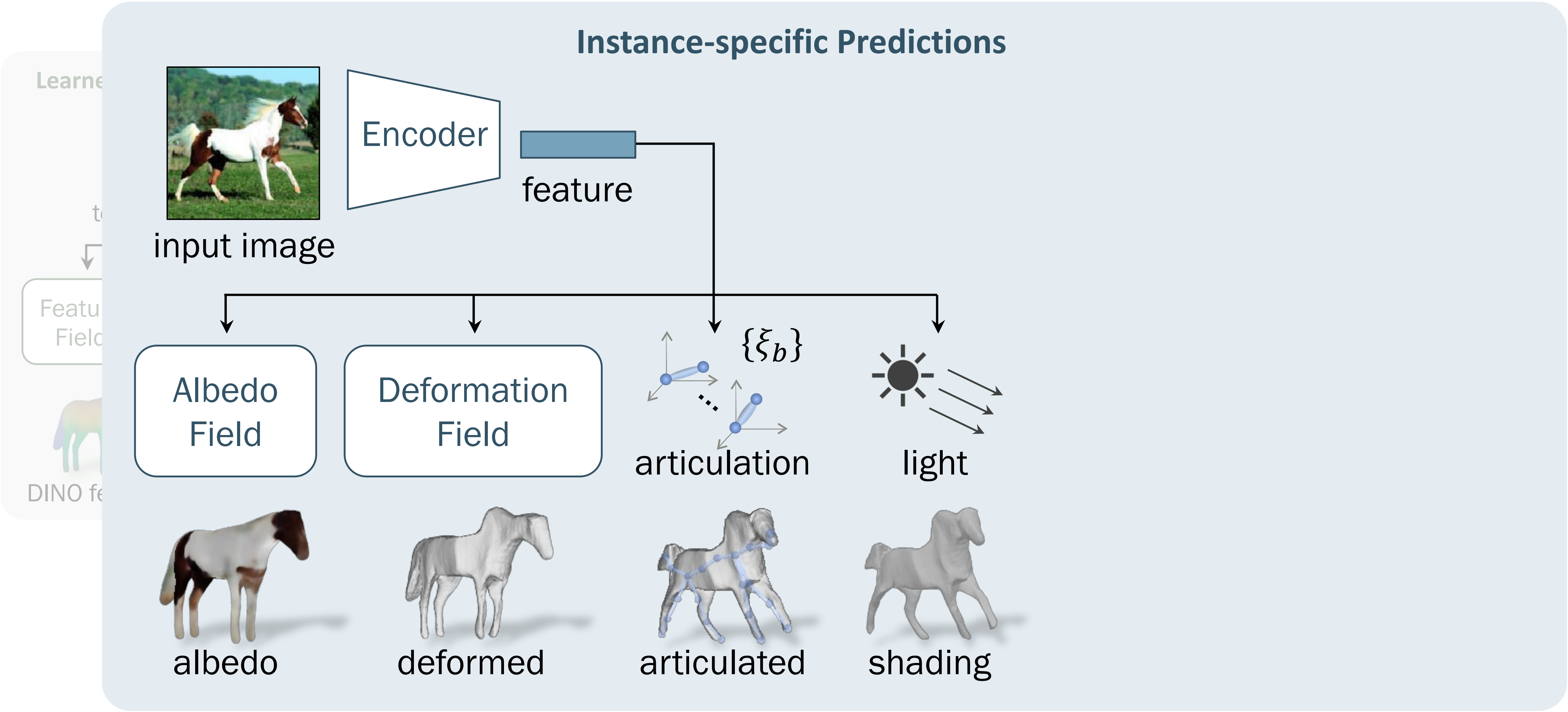
Mesh
✓ Easy to render
✓ Easy to articulate

\+

DMTet
✓ Differentiable
✓ Regular (no self-intersection)

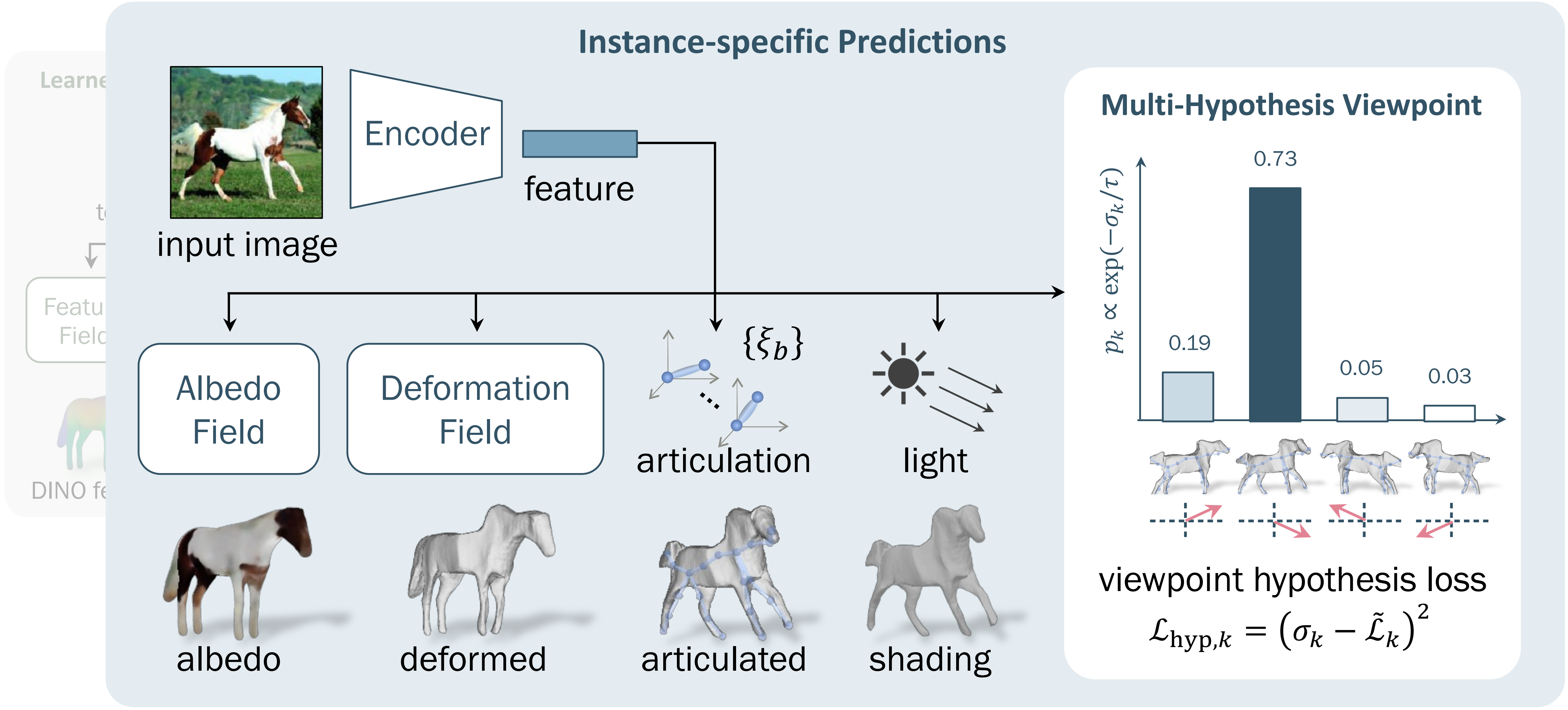
[1] Deep Marching Tetrahedra: a Hybrid Representation for High-Resolution 3D Shape Synthesis. Shen et. al. NeurIPS 2021.

Hierarchical Shape Prediction



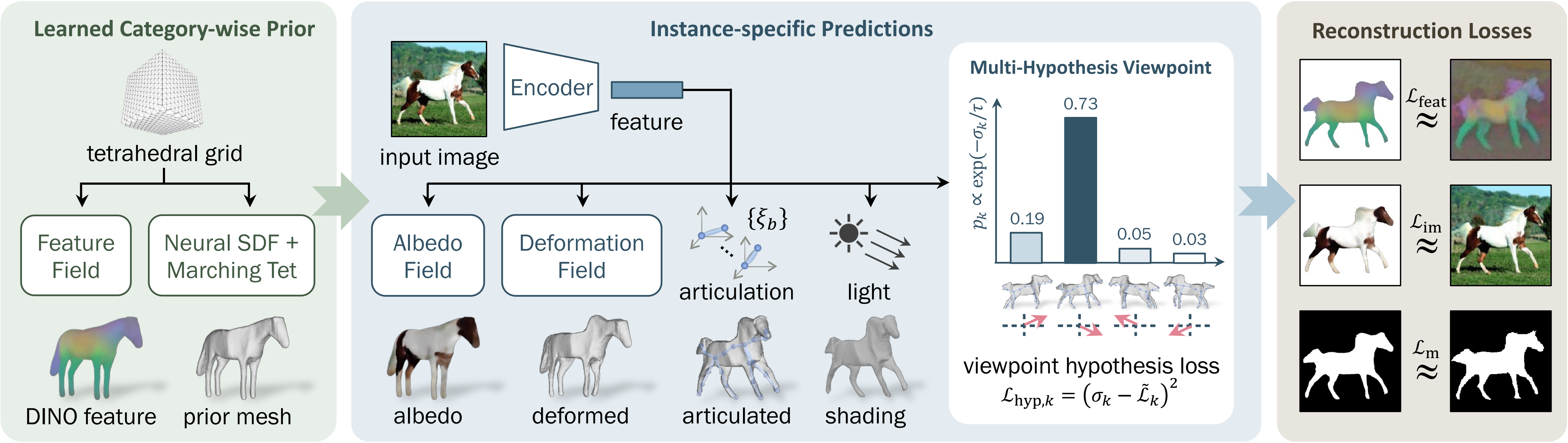
[1] Deep Marching Tetrahedra: a Hybrid Representation for High-Resolution 3D Shape Synthesis. Shen et. al. NeurIPS 2021.

Hierarchical Shape Prediction



[1] Deep Marching Tetrahedra: a Hybrid Representation for High-Resolution 3D Shape Synthesis. Shen et. al. NeurIPS 2021.

End-to-End Training with Image Rendering Losses

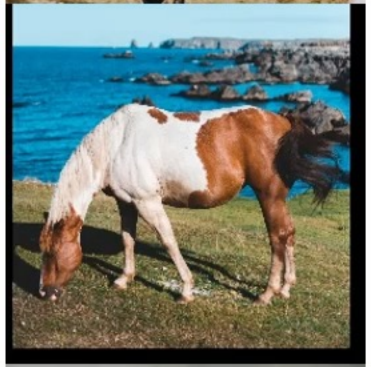


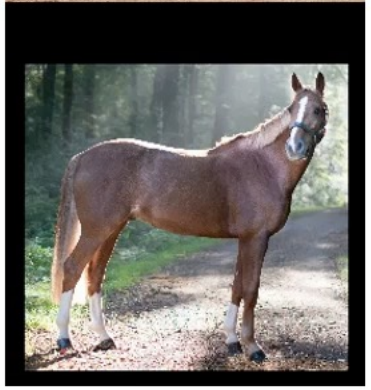
[1] Deep Marching Tetrahedra: a Hybrid Representation for High-Resolution 3D Shape Synthesis. Shen et. al. NeurIPS 2021.
[2] Emerging Properties in Self-supervised Vision Transformers. Caron et. al. ICCV 2021.







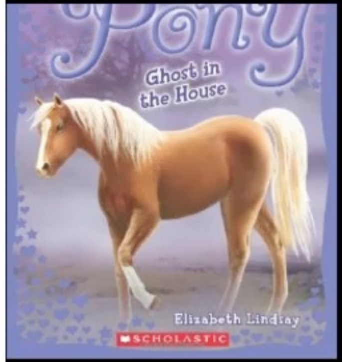






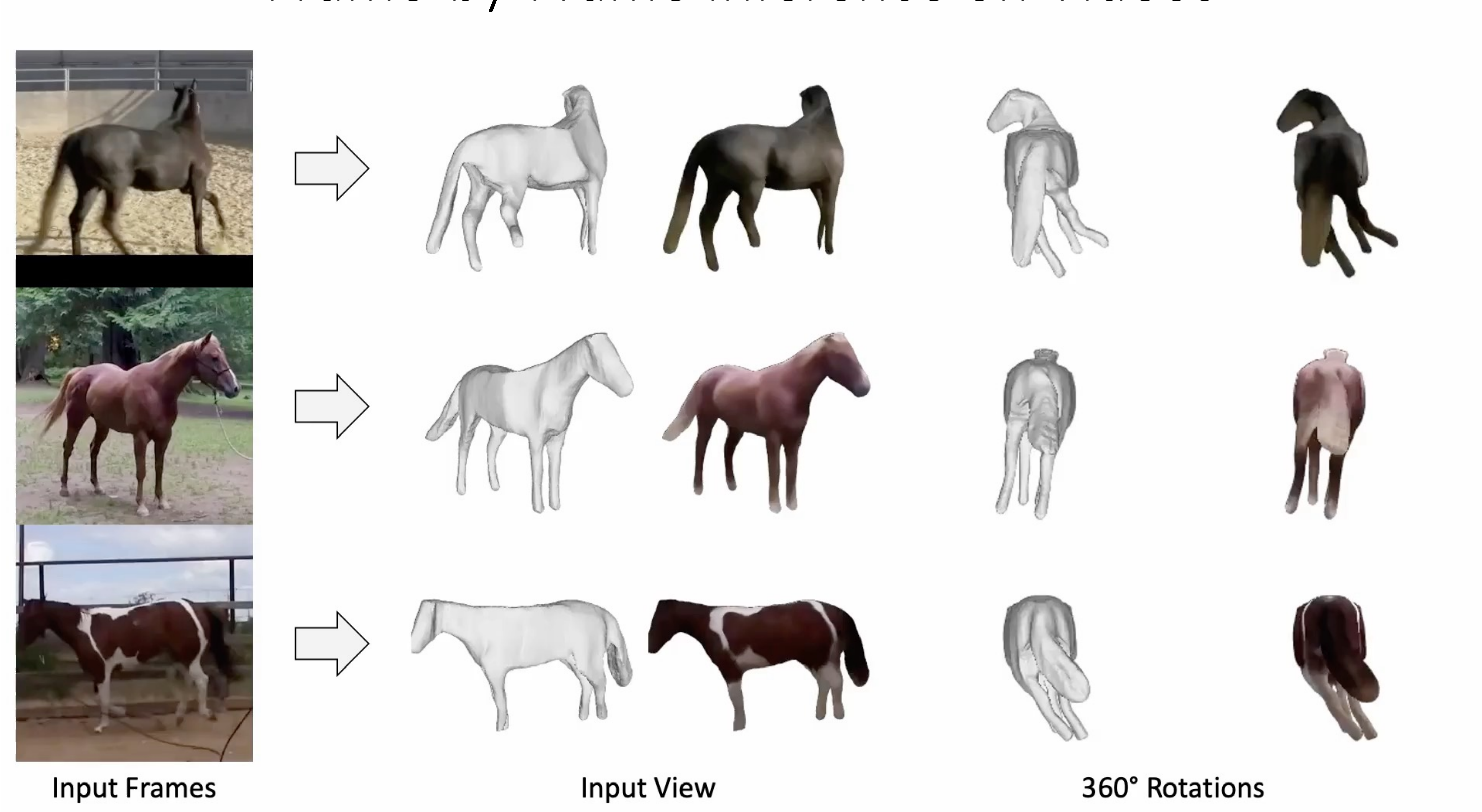






Ghost in
the House
Elizabeth Lindsay
SCHOLASTIC



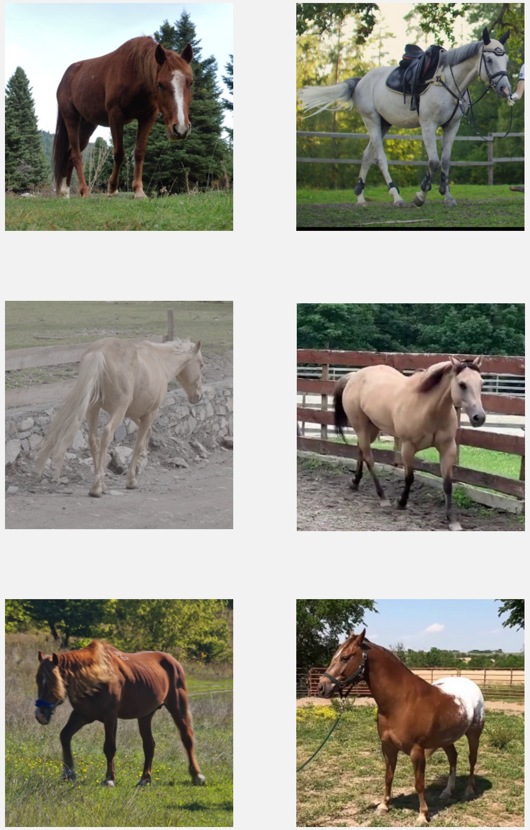

Frame-by-Frame Inference on Videos



3D Printed Horse Reconstruction

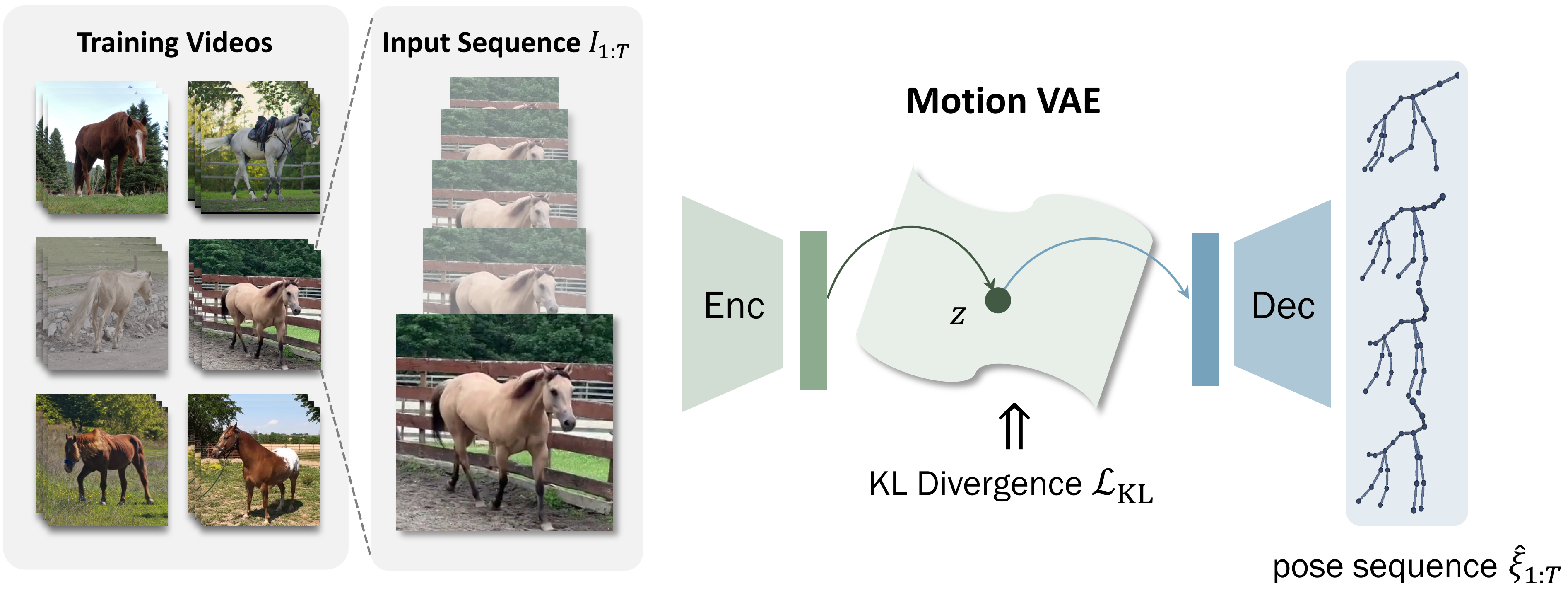
Learning Articulated 3D Motion Prior

Training Images

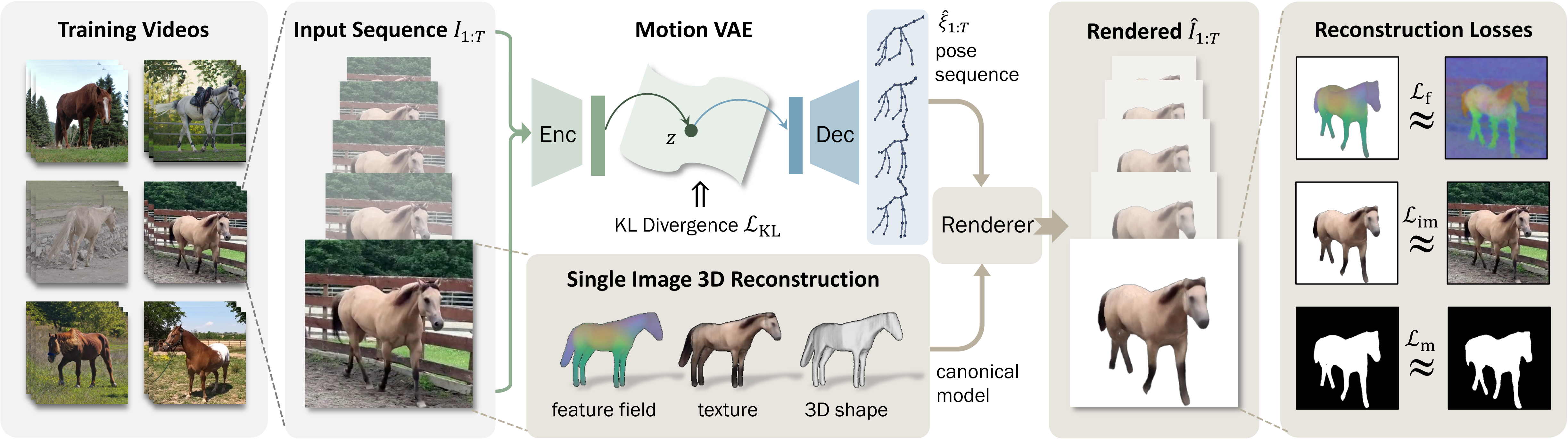


Learning Articulated 3D Motion Prior

Learning Articulated 3D Motion Prior

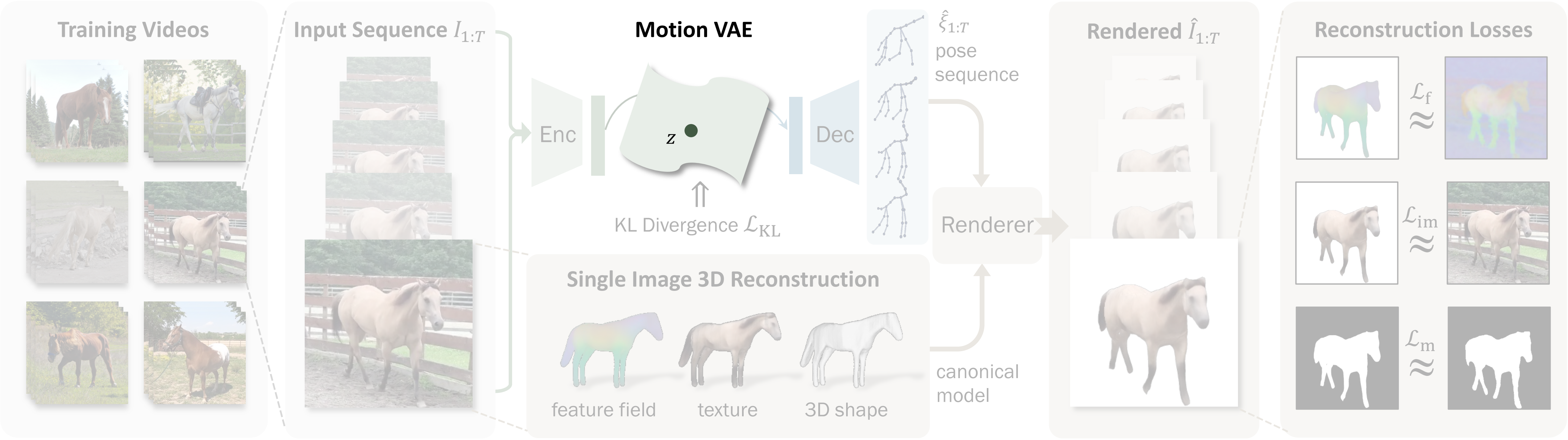


Learning Articulated 3D Motion Prior



Trained with 2D reconstruction losses only without any pose annotations!

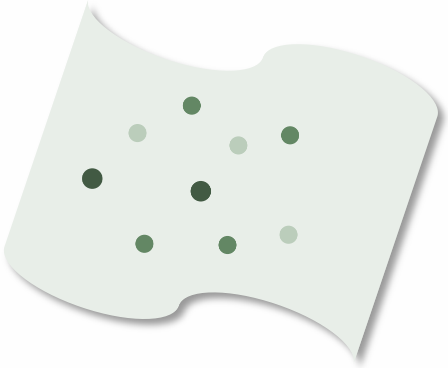
Learning Articulated 3D Motion Prior



Trained with 2D reconstruction losses only without any pose annotations!

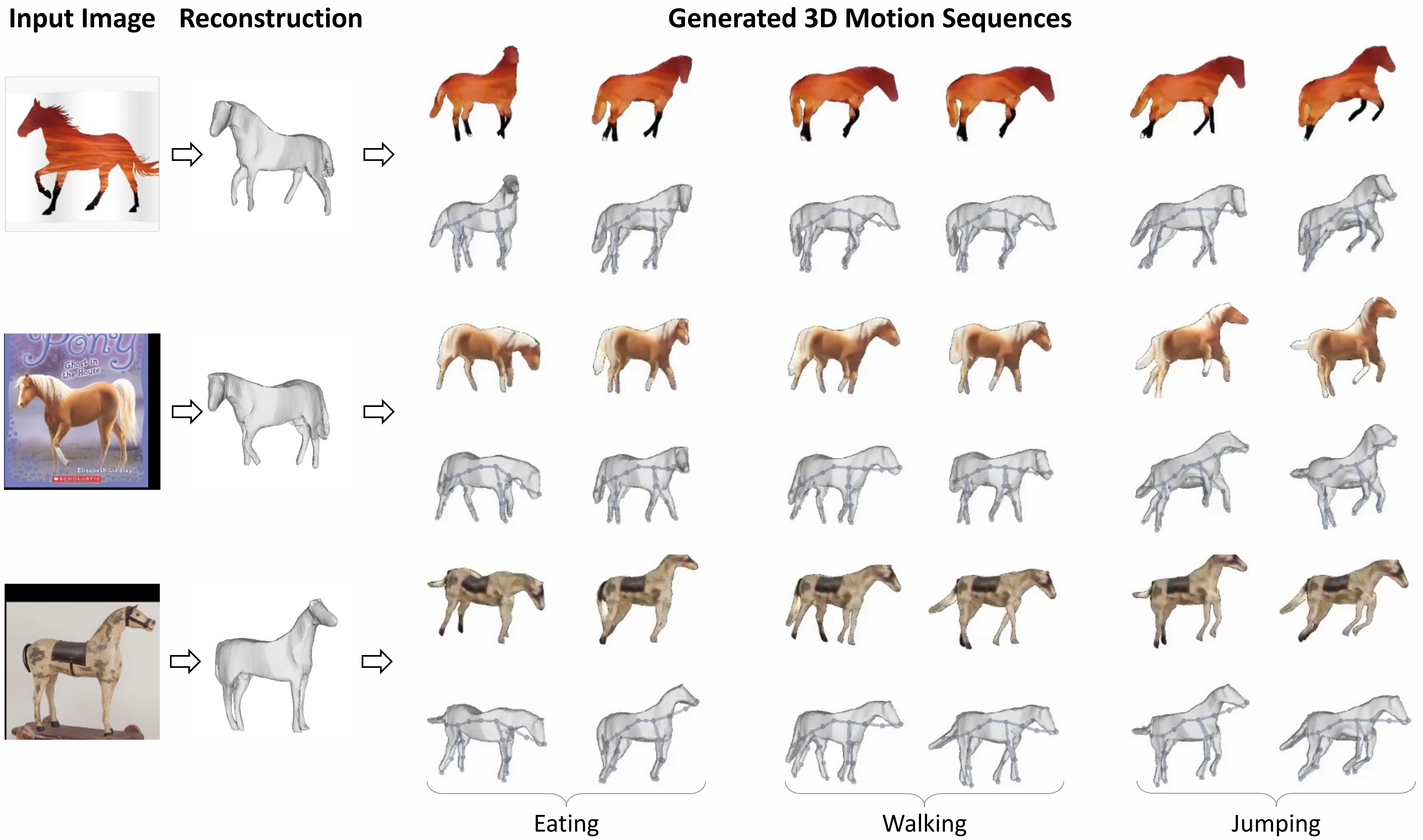
Learning Articulated 3D Motion Prior

Motion VAE



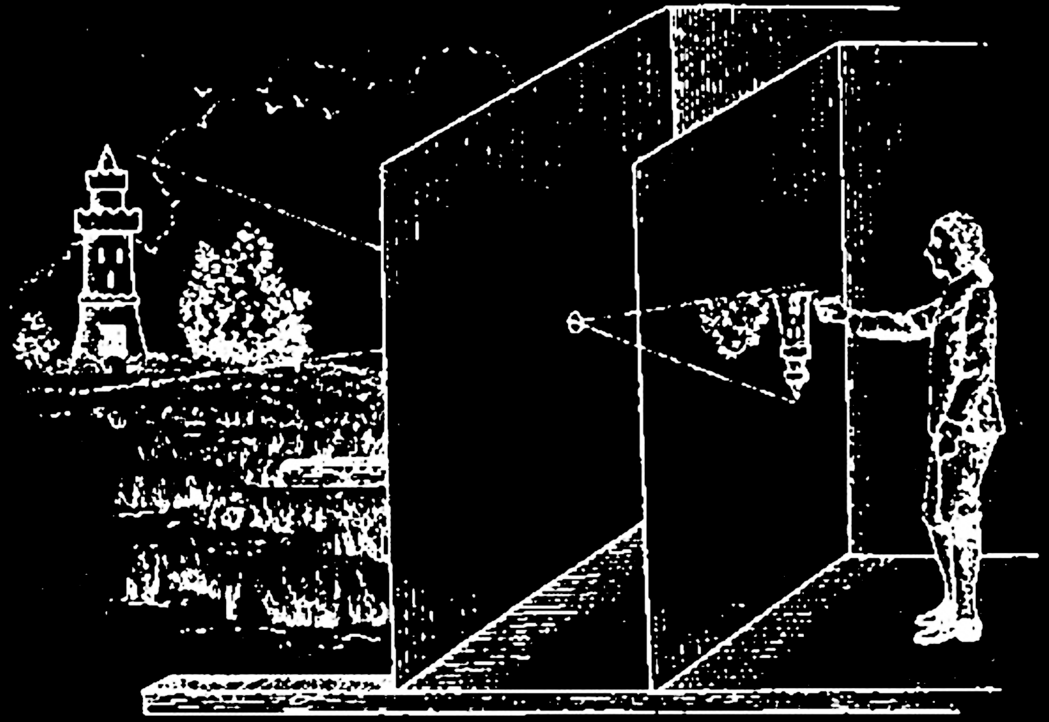
Random Samples

Generated 3D Motion Sequences



Input Image
Reconstruction
Generated 3D Motion Sequences
Pony
Ghost in the House
Elisabeth Lindsay
SCHOLASTIC
Eating
Walking
Jumping

It's a ~~3D~~ World, After All
Physical



Physics is the key to interpretability and generality!

Learning Dynamic 3D Objects in the Wild

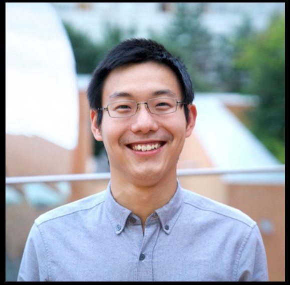
Stanford University

Shangzhe Wu

PostDoc at Stanford SVL

UNIVERSITY OF OXFORD

Amazing Advisors & Collaborators



Jiajun Wu

Andrea Vedaldi

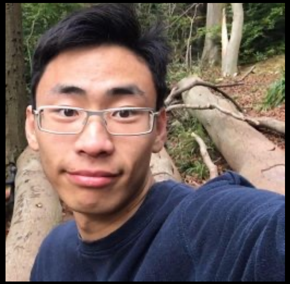
Christian Rupprecht

Noah Snavely

Yunzhi Zhang

Tomas Jakab

Ruining Li



Zirui Wang

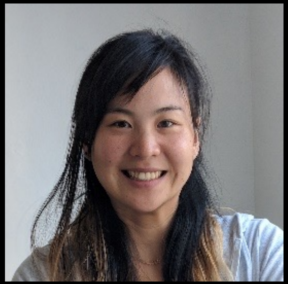
Felix Wimbauer

Keqiang Sun

Hongsheng Li

Ameesh Makadia

Richard Tucker



Angjoo Kanazawa