[CVPR 2025]

AniMer: Animal Pose and Shape Estimation Using Family Aware Transformer

Jin Lyu[1,*], Tianyi Zhu[2,*], Yi Gu[3], Li Lin[1,4], Pujin Cheng[1,4], Yebin Liu[5], Xiaoying Tang[1,†], Liang An[5,†]
[1]Southern University of Science and Technology
[2]China Mobile Communications Company Limited Research Institute
[3]The Hong Kong University of Science and Technology
[4]The University of Hong Kong [5]Tsinghua University

Presenter: Gyeongsu Cho
Fri May 15, 2025

Contents

- Introduction
- Method
- Experiments
- Conclusion

Introduction

Animer

AniMer is to estimate the pose and shape from a single image across quadrupedal species

Motivation for animal pose and shape estimation

- Graphics, VR/AR, and Robotics
  - Enables creation of **realistic animal avatars**
  - Enhances robot **perception systems** for animal-rich environments

- Scientific Research & Behavioral Analysis
  - Quantitative study of animal behavioral, posture, and interaction
  - Advances research in **zoology**, **ethology**, and **neuroscience**

Challenge: Multi-species variability, lack of datasets

[CVPR 2024] BITE

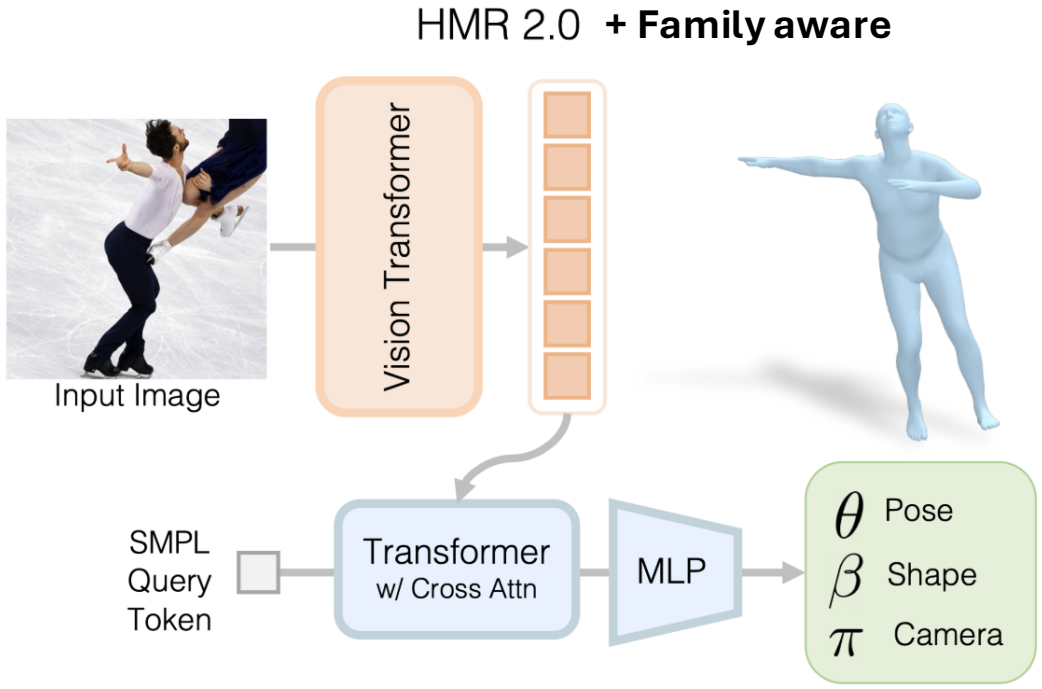
[ACCV 2024] Dessie

[ICCV 2023] Animal3D dataset

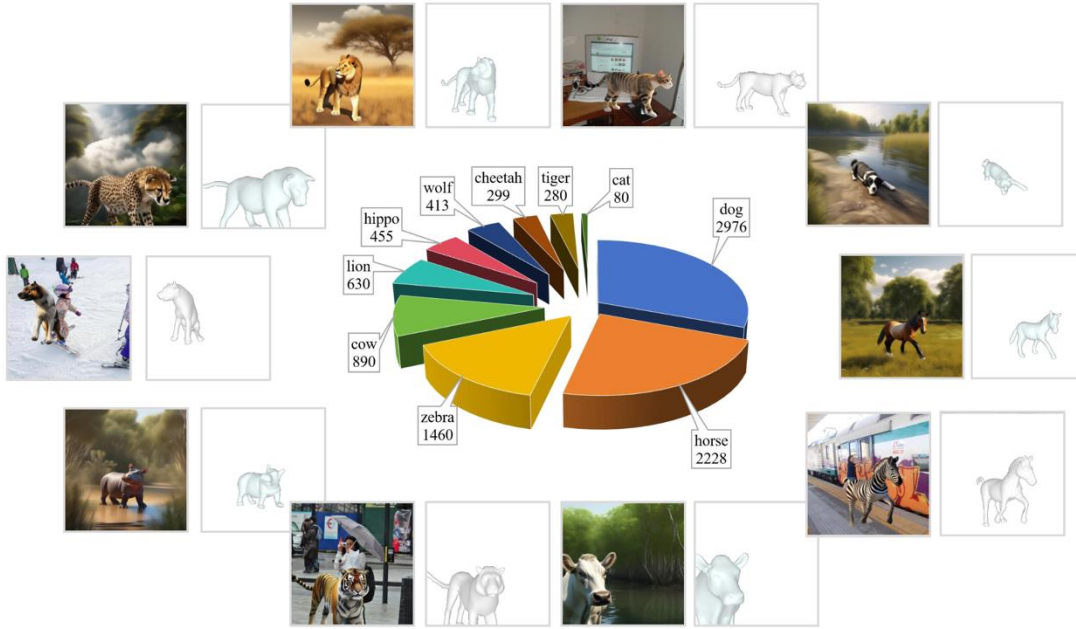
Most model focus on single species like dogs or horses.

Multi-species datasets like Animal3D remain limited in diversity and generalization.

Key Idea



Family-aware Transformer



CtrlAni3D dataset

1. Utilizing quadruped family information for pose and shape estimation
2. A large-scale synthetic dataset of diverse quadrupeds (10 species, 9.7K images)

Preliminaries: SMAL

$\beta \in \mathbb{R}^{41}$

$\theta \in \mathbb{R}^{35\times 3}$

SMAL $\mathcal{M}(\beta, \theta, \gamma)$

$W \in \mathbb{R}^{35\times 3889}$

$J \in \mathbb{R}^{35\times 3}$

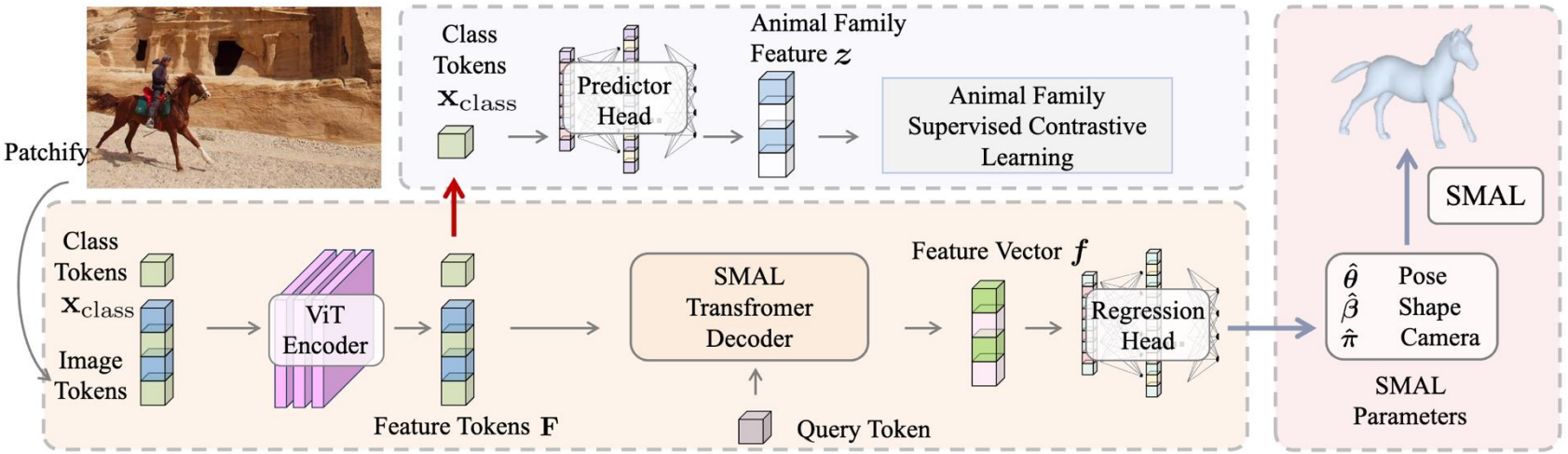
$V \in \mathbb{R}^{3889\times 3}$ $F \in \mathbb{N}^{7774\times 3}$

[CVPR 2017] SMAL

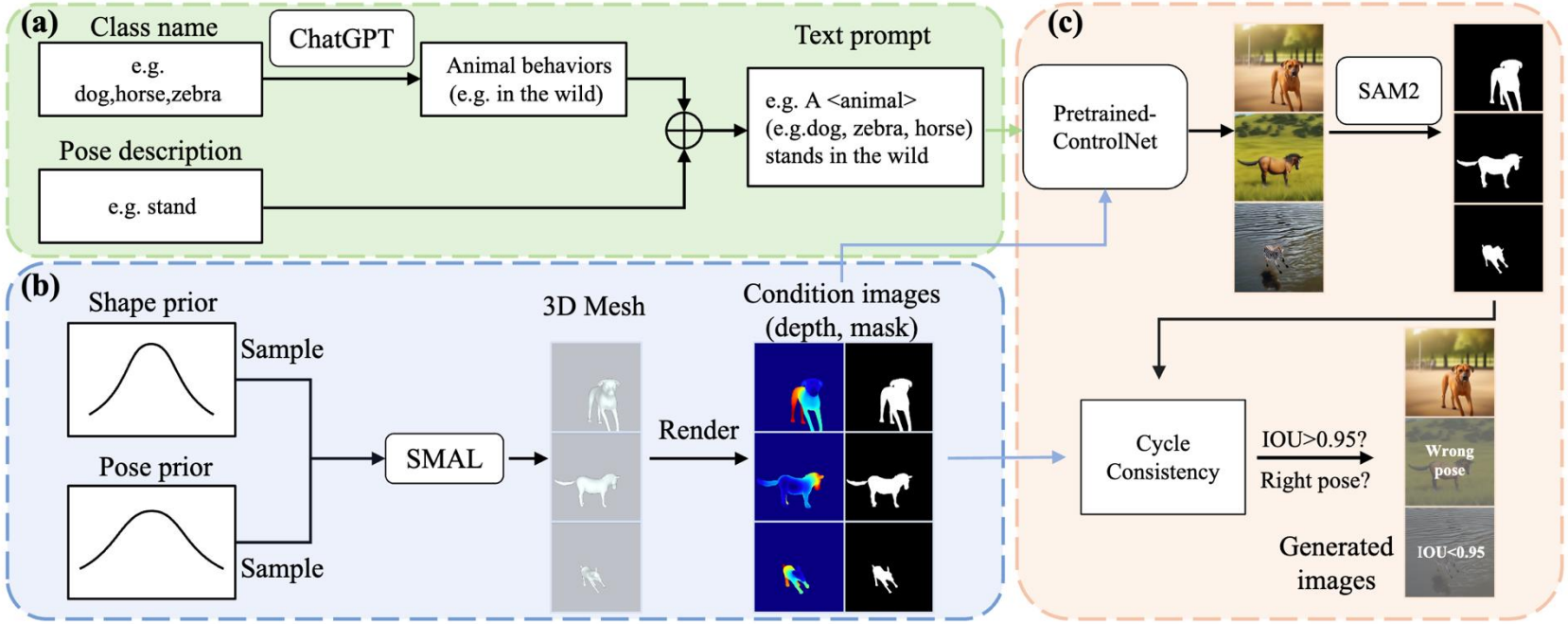
SMAL is a 3D parametric shape model designed specifically for representing the shape and pose of quadruped animals

Method

Overview of AniMer

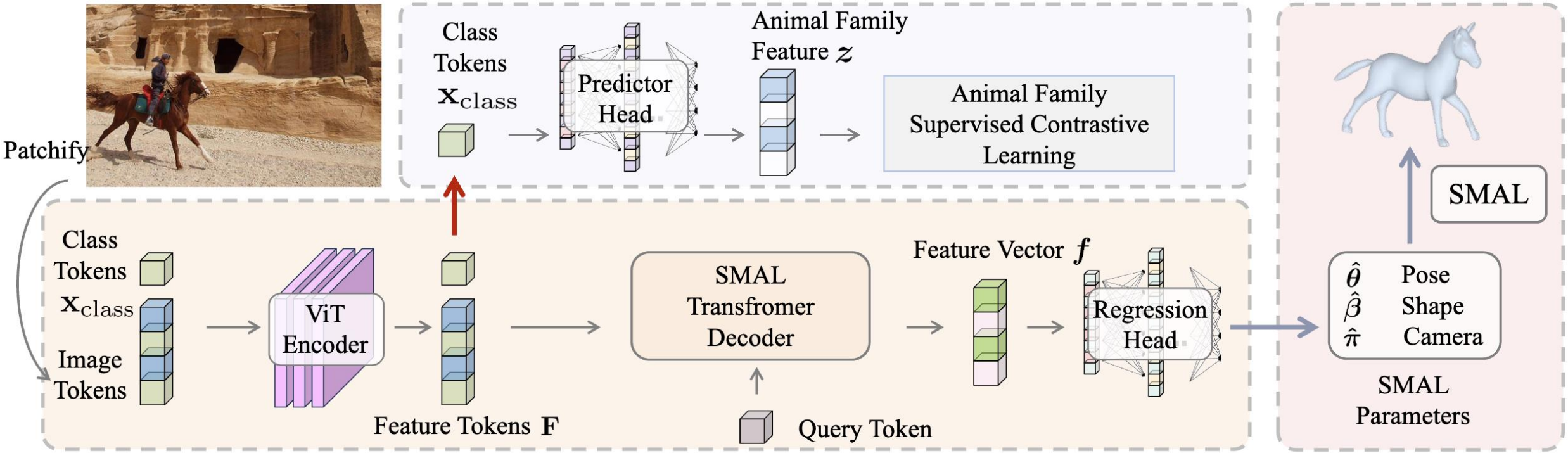


AniMer architecture



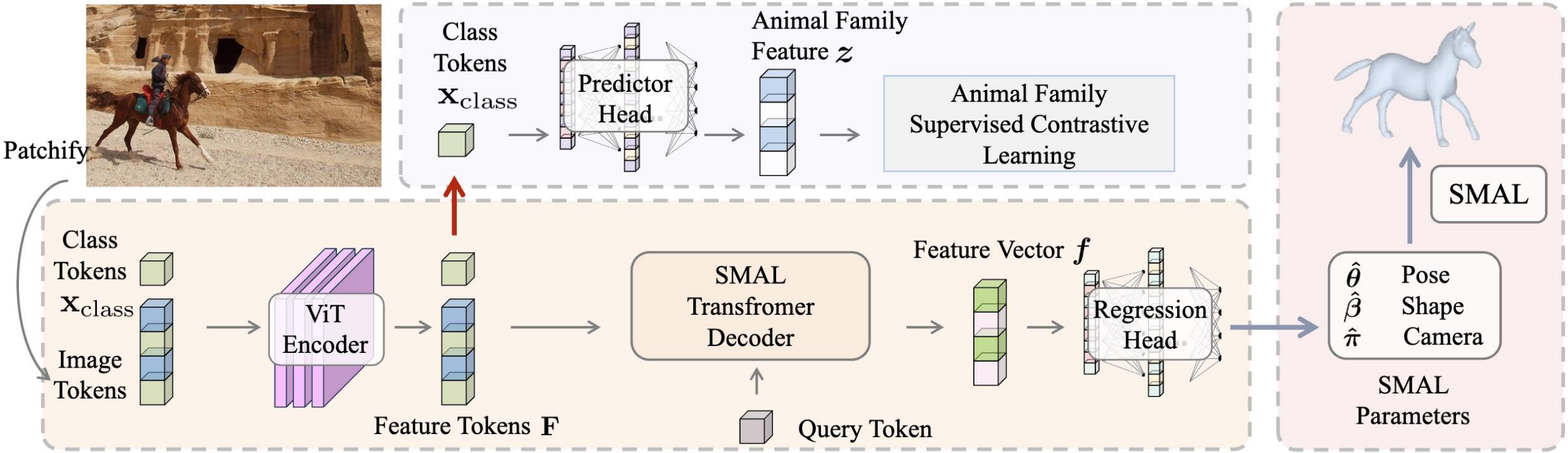
AniMer data generation pipeline

Architecture of AniMer



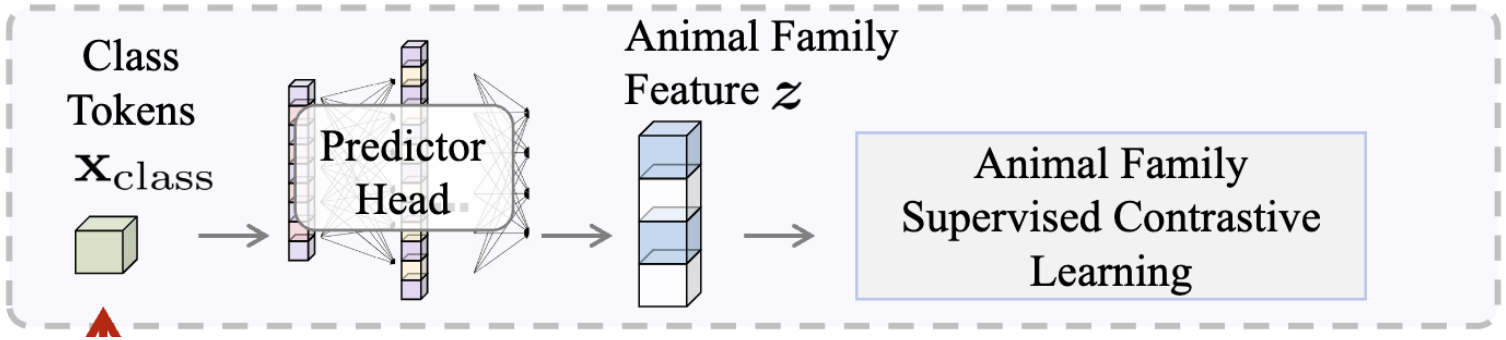
- ViT encoder processes the input image as patch tokens for rich visual representation.
- Transformer decoder refines features for two tasks: SMAL parameter regression and family-aware contrastive learning.

Architecture of AniMer



- ViT encoder is initialized with pretrained weights.
- AniMer is trained in two stages.

Family Aware Contrastive Learning



Contrastive loss clusters family-specific features using supervised contrastive learning.

$$\mathcal{L}_{\text{con}} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \frac{\exp\left(\boldsymbol{z_i} \cdot \boldsymbol{z_p}\right)}{\sum_{o \in O(i)} \exp\left(\boldsymbol{z_i} \cdot \boldsymbol{z_o}\right)}.$$

- **Class token** represents animal family identity during feature extraction.
- **Predictor head** generates a family-specific feature vector z.

Training

Total Loss combines multiple objectives:

$$\mathcal{L}_{\text{total}} = \lambda_{3D}\mathcal{L}_{3D} + \lambda_{2D}\mathcal{L}_{2D} + \lambda_{\text{prior}}\mathcal{L}_{\text{prior}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}.$$

3D Loss penalizes shape (β), pose (θ), and 3D joint errors:

$$\mathcal{L}_{3D} = \lambda_\beta ||\hat{\beta} - \beta||_2^2 + \lambda_\theta ||\hat{\theta} - \theta||_2^2 + ||\hat{K}_{3D} - K_{3D}||_1,$$

2D projection loss supervises projected 3D keypoints against ground truth 2D annotations.

$$\mathcal{L}_{2D} = ||\pi(\hat{K}_{3D}) - K_{2D}||_1$$

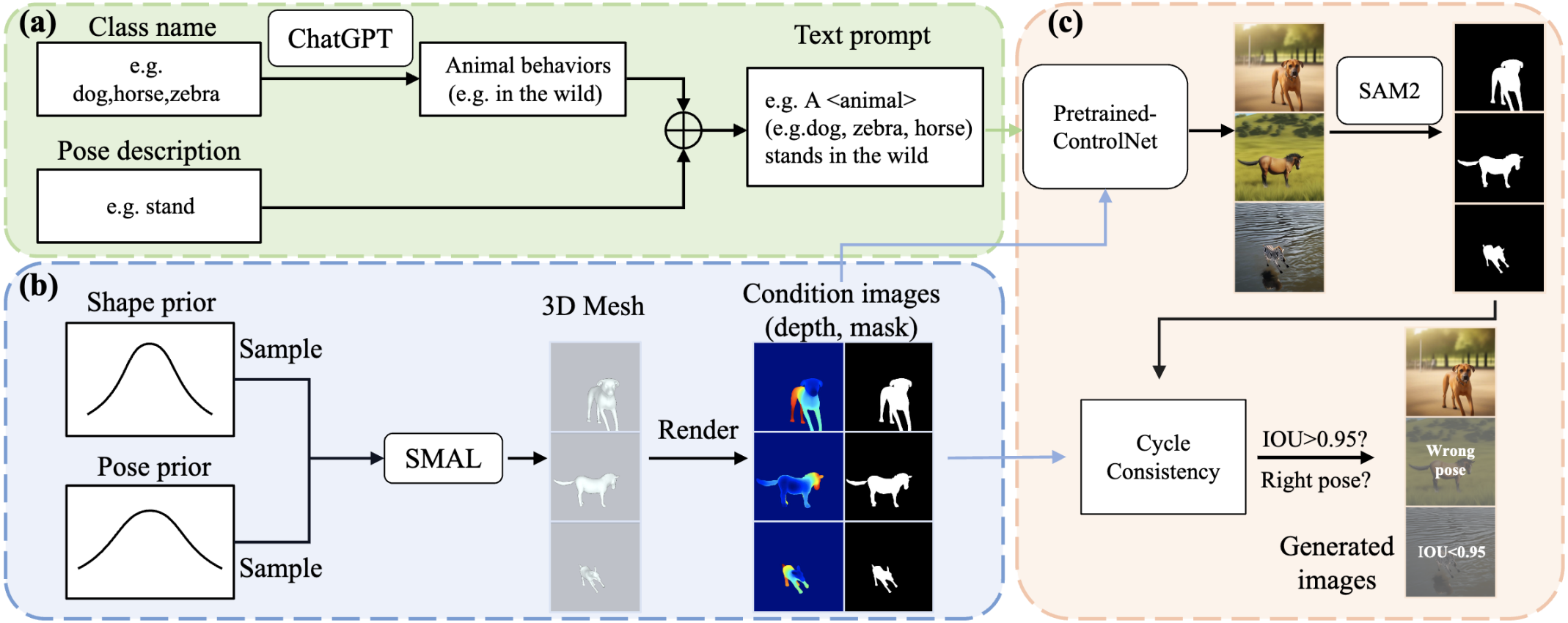
Prior loss enforces predicted parameters to stay close to a prior distribution.

$$\mathcal{L}_{\text{prior}} = \lambda_\beta(\hat{\beta}-\mu_\beta)^T\Sigma_\beta^{-1}(\hat{\beta}-\mu_\beta)+(\hat{\theta}-\mu_\theta)^T\Sigma_\theta^{-1}(\hat{\theta}-\mu_\theta),$$

Adversarial loss uses a discriminator to encourage realistic pose and shape outputs

$$\mathcal{L}_{\text{adv}} = \sum_k (D_k(\theta, \beta) - 1)^2$$

Dataset generation pipeline of AniMer



AniMer generates **CtrlAni3D** dataset by combining text prompts from animal classes and poses with SMAL-based 3D conditioning

CtrlAni3D Dataset

SMAL Structure Condition

- Sampled shape and pose parameters are used to generate 3D meshes, rendered into depth maps and masks.

Text condition

- Text prompts describing animal behavior are generated using species names and pose labels, completed with ChatGPT.

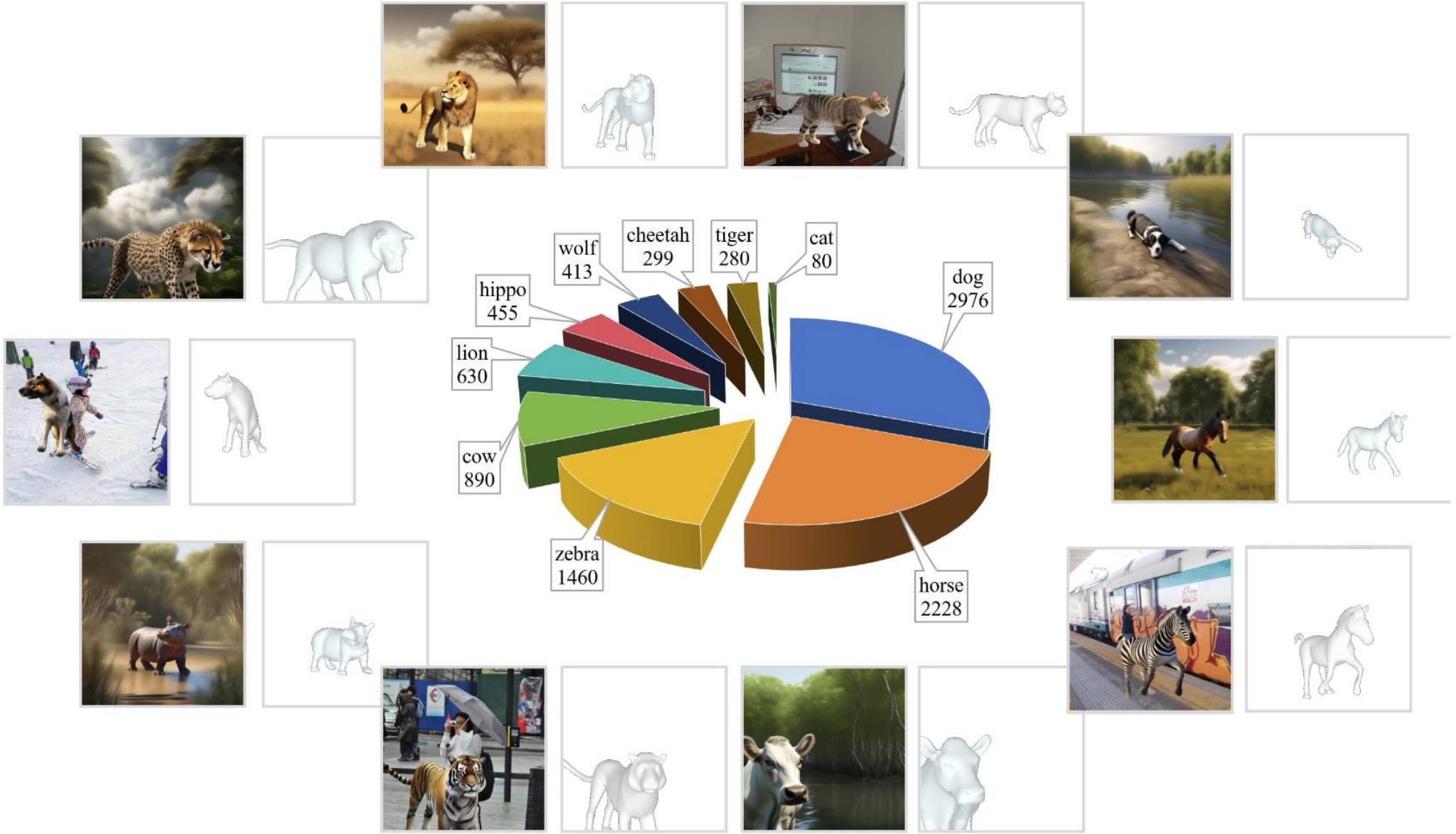
Semi-Automated Filtering

- SAM2 and cycle-consistency check are used to validate alignment between generated images and mesh conditions.

Backgrounds

- Natural scene diversity is enhanced by combining AI-generated and COCO-sampled backgrounds during image synthesis.

CtrlAni3D Dataset



CtrlAni3D dataset statistics and visual samples

Experiments

Experiments

Dataset

- Training on a combined dataset of 41.3k images from 2D and 3D sources: Animal3D, CtrlAni3D, Animal Pose, APT-36K, AwA-Pose, Stanford Extra, Zebra Synthetic.

Baseline

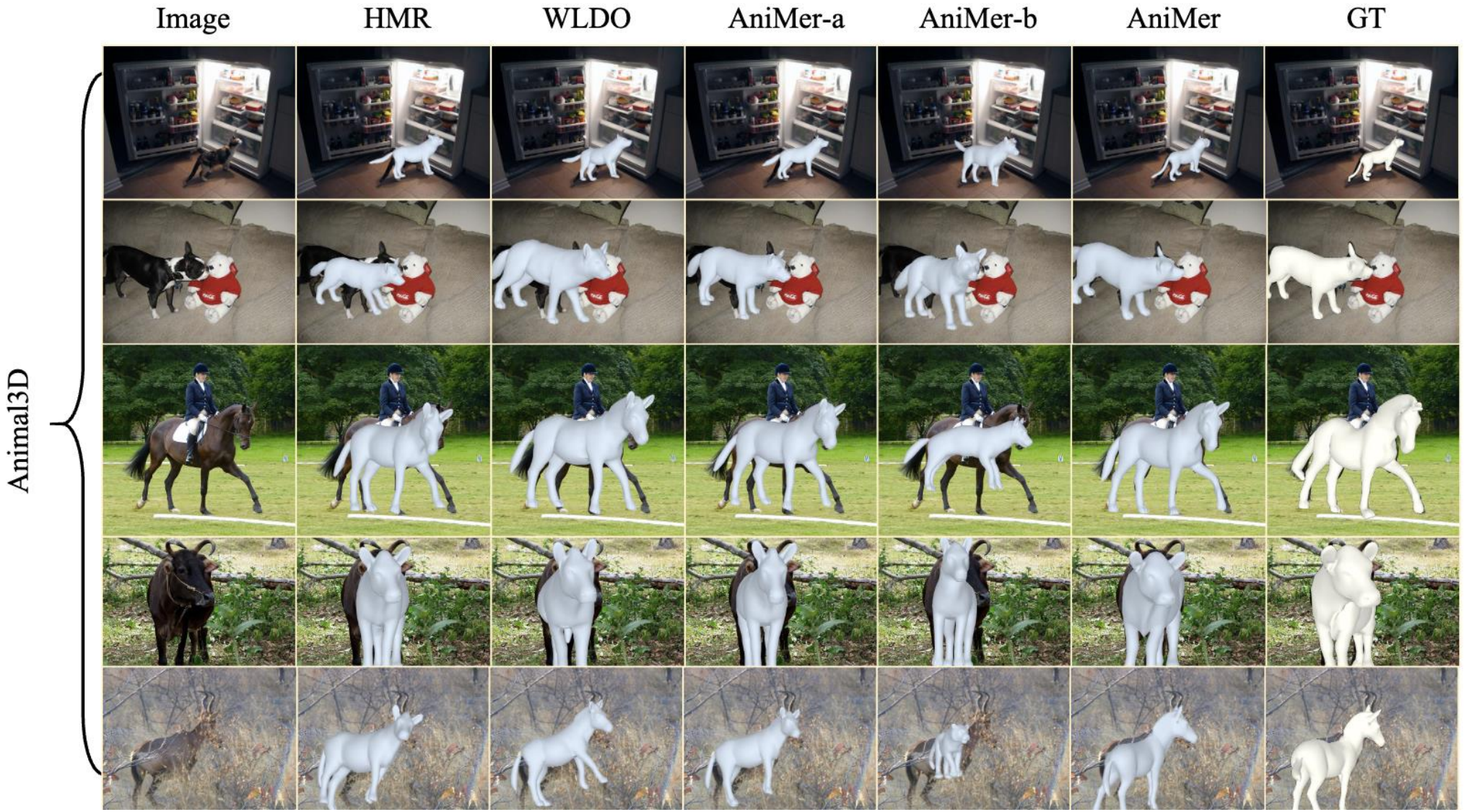
- Compared against [CVPR 2018] HMR, [ECCV 2020] WLDO, and [ICCV 2023] HMR2.0.

Evaluation metrics

- 3D: PA-MPJPE (joints), PA-MPVPE (mesh).
  2D: PCK@0.1, PCK@0.15, AUC.

Experiments

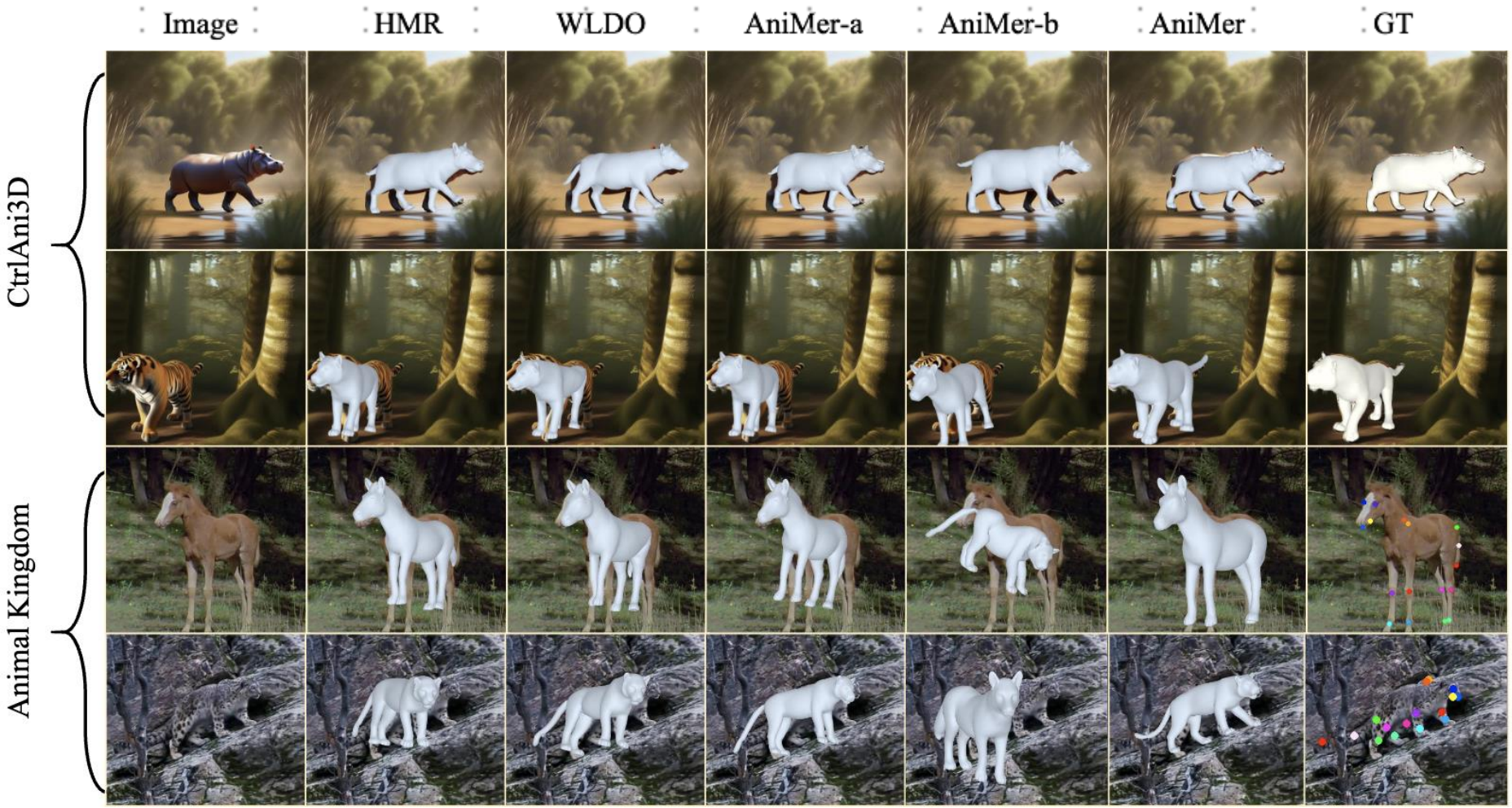
AniMer-a: replaces ViT backbone with ResNet152.
AniMer-b: discards ViT pretraining.



Qualitative comparisons of AniMer

Experiments

AniMer-a: replaces ViT backbone with ResNet152.
AniMer-b: discards ViT pretraining.



Qualitative comparisons of AniMer

Experiments

AniMer-a: replaces ViT backbone with ResNet152.
AniMer-b: discards ViT pretraining.

Dataset	Animal3D				CtrlAni3D				Animal Kingdom			
Metric	AUC↑	P@H↑	PAJ↓	PAV↓	AUC↑	P@H↑	PAJ↓	PAV↓	AUC↑	P@H↑	P@0.1↑	P@0.15↑
HMR	76.3	60.8	123.5	133.9	80.8	67.0	123.5	133.9	70.2	64.0	12.8	25.6
WLDO	78.2	68.7	112.3	125.2	88.7	86.7	71.5	83.4	70.1	64.3	14.6	27.6
AniMer-a	75.2	57.2	115.5	128.7	80.3	66.0	117.0	129.4	68.9	62.5	10.2	21.3
AniMer-b	60.6	38.9	147.9	157.6	78.5	65.9	102.3	112.6	45.4	31.8	4.0	9.2
AniMer	**88.9**	**89.5**	**80.4**	**85.7**	**93.8**	**95.4**	**44.1**	**47.6**	**82.9**	**83.7**	**34.9**	**54.7**

Qualitative comparisons of AniMer

Experiments

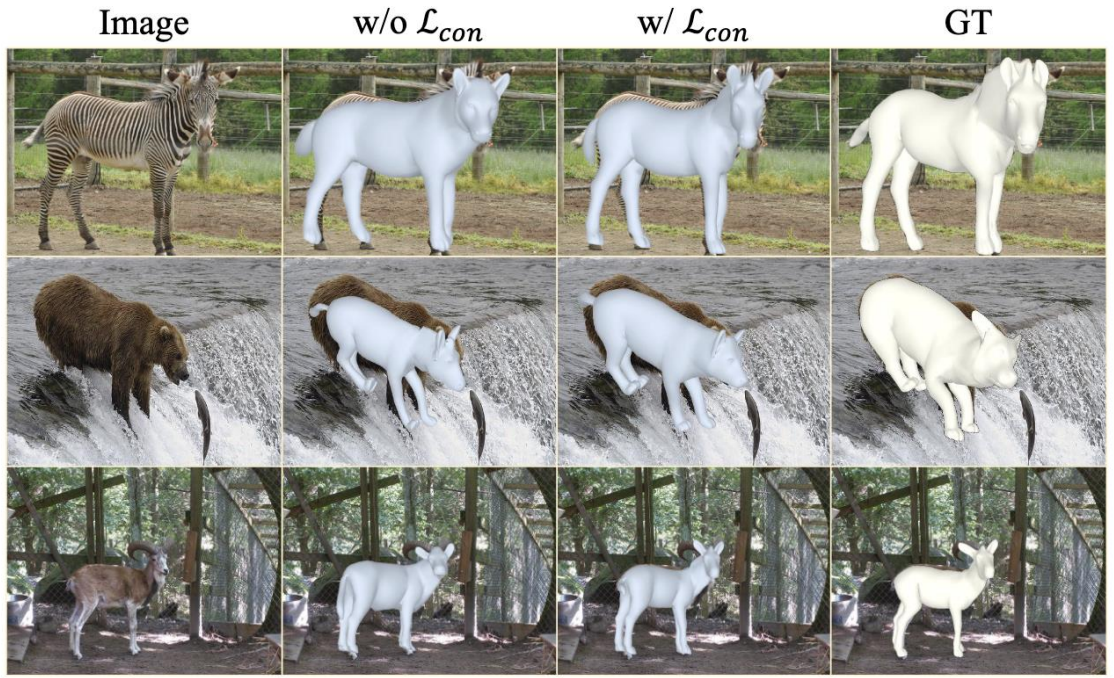
AniMer-c: trains the AniMer model for one stage.

Dataset	Animal3D				CtrlAni3D				Animal Kingdom			
Metric	AUC↑	P@H↑	PAJ↓	PAV↓	AUC↑	P@H↑	PAJ↓	PAV↓	AUC↑	P@H↑	P@0.1↑	P@0.15↑
HMR2.0	86.7	84.6	94.1	98.5	91.8	93.0	60.9	66.4	77.3	73.9	22.7	40.2
AniMer-c	87.2	86.3	85.9	90.4	91.7	93.4	59.5	64.2	80.6	80.4	28.6	47.5
AniMer	**88.9**	**89.5**	**80.4**	**85.7**	**93.8**	**95.4**	**44.1**	**47.6**	**82.9**	**83.7**	**34.9**	**54.7**

Qualitative comparisons of AniMer

Experiments

Method	Animal3D		CtrlAni3D		Animal Kingdom	
	PAJ↓	PAV↓	PAJ↓	PAV↓	AUC↑	P@0.1↑
w/o $\mathcal{L}_{\text{con}}$*	82.5	88.0	54.6	59.2	81.4	30.4
w/ $\mathcal{L}_{\text{con}}$*	**80.8**	**86.0**	**46.1**	**50.2**	**81.9**	**32.1**
w/o $\mathcal{L}_{\text{con}}$	81.3	86.7	44.7	48.4	82.7	34.4
w/ $\mathcal{L}_{\text{con}}$	**80.4**	**85.7**	**44.1**	**47.6**	**82.9**	**34.9**



Ablation study of AniMer

Experiments

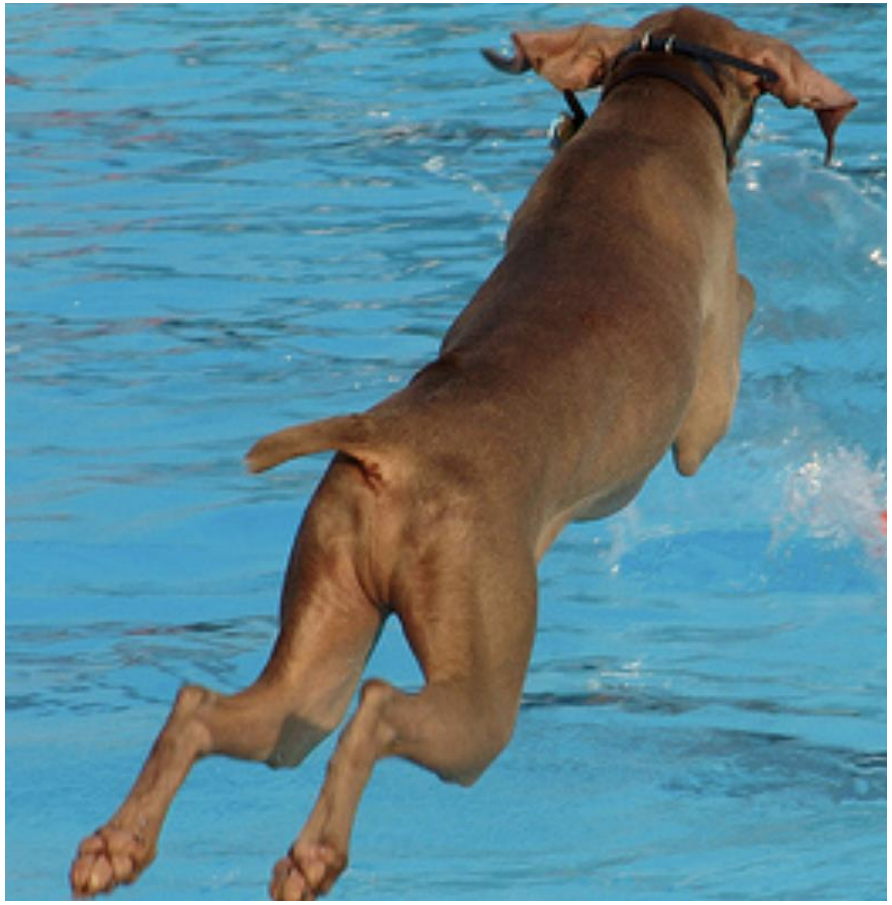
Limitations of LGM

Conclusion

Conclusion

AniMer is to estimate the pose and shape from a single image across quadrupedal species

Conclusion



Input image

Mesh overlay of AniMer

AniMer is to estimate the pose and shape from a single image across quadrupedal species

Appendix

Metrics

PA-MPJPE (Procrustes Aligned Mean Per Joint Position Error)

PA-MPVPE (Procrustes Aligned Mean Per Vertex Position Error)

PCK@HTH (Percentage of Correct Keypoints at Head Top Height)