DualPM: Dual Posed-Canonical Point Maps for 3D Shape and Pose Reconstruction

Supplementary Material

Figure A1. **Results on unseen categories.** A version of our model trained only on the horse category also demonstrates robust generalization to the unseen categories such as cow and sheep, despite being trained solely with a single horse model.

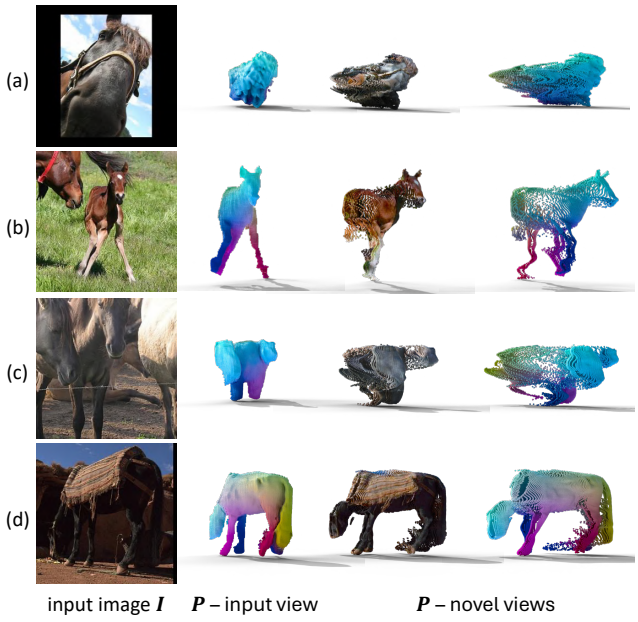

Figure A2. **Typical failure cases.** We illustrate representative failure cases caused by (a) extreme viewpoints, (b) shapes and poses far from the training distribution, and (c-d) inaccuracies in the object segmentation masks.

Appendix

A. Generalization to unseen categories

Given the generalization capabilities of our method demonstrated within a single category, we analyze the generalization of a model trained on a single category to unseen categories. Specifically, we consider a model trained on horses and evaluate its performance on cow and sheep categories. We evaluate our approach using the same datasets as for the horses, PASCAL VOC [1] and Animodel [2], following the same evaluation protocol, with results reported in Tab. A1. Furthermore, we provide additional qualitative results on the same dataset in Fig. A1. Our method exhibits strong zero-shot generalization to these categories, outperforming state-of-the-art approaches on both datasets, despite being trained exclusively on a single horse model.

B. Limitations

Despite demonstrating surprising generalization, a current limitation of our method is that any additional synthetic 3D models added to the training dataset would have to be in the same canonical space as the training data. Addressing the challenging problem of aligning the canonical spaces of multiple 3D models would allows us to train on significantly larger datasets which could in turn lead to significant gains in performance for our proposed model. Another limitation of our method is that it is not specifically trained to handle occlusions caused by other objects. This is a limitation shared with other methods, such as 3D-Fauna [6, 11]. To address this, we plan to extend the data generation pipeline to include synthetic occlusions. Additionally, as the 3D reconstruction problem is often ambiguous for the unseen parts of objects, our method predicts only the expectation over

Method	PCK (%)		Chamfer Distance (cm)			
			Real-Sized		Normalized	
	Cow	Sheep	Cow	Sheep	Cow	Sheep
A-CSM [5]	26.3	28.6	6.71 ± 1.81	2.84 ± 0.77	2.35 ± 0.68	2.48 ± 0.70
MagicPony [11]	42.5	41.2	7.22 ± 1.53	3.43 ± 0.73	2.53 ± 0.59	3.00 ± 0.68
Farm3D [2]	40.2	36.1	6.91 ± 1.49	3.79 ± 0.55	2.41 ± 0.54	3.31 ± 0.49
3D-Fauna [6]	—	—	9.19 ± 2.40	3.51 ± 0.88	3.20 ± 0.80	3.06 ± 0.76
Ours	**63.0**	**64.2**	**4.74** ± 1.40	**2.32** ± 0.78	**1.67** ± 0.55	**2.03** ± 0.71

Table A1. **Evaluation on unseen cow and sheep categories.** We evaluate on PASCAL VOS, reporting PCK@0.1 (higher is better ↑), and on Animodel [2], reporting the bi-directional Chamfer Distance in centimeters (lower is better ↓). Our model, trained solely data from a single horse model, outperforms state-of-the-art approaches, which were trained on data that included these specific categories.

all possible reconstructions, which can lead to unrealistic results for the invisible regions. We illustrate our typical failure cases in Fig. A2.

C. Technical details

Network architecture. We obtain the segmentation mask M using the Segment Anything method [4]. The feature extractor Ψ is based on [12] which combines pre-trained DINOv2 [7] and StableDiffusion [8] networks. Training image features are reduced to a 64-dimensional space using PCA following [10]. The dual point map predictors Φ_Q and Φ_P leverage a convolutional U-Net architecture based on [9], comprising two blocks each and trained from scratch. We predict $N = 4$ layers for layered amodal point maps as more have little effect on the performance (**??**), likely due to the low frequency of multiple self-occlusions in our datasets. The number of layers can be easily increased should the data require it. The output resolution of the layered point maps is set to 160×160.

Training. We use the Adam optimizer [3] for training. Our model is trained for 100k steps with a batch size of 12. The learning rate is set to 6×10^{-4}, with a step scheduler applied, featuring a 30k-step period and a decay factor of 0.5.

Training dataset. The training dataset consists of approximately 30k rendered images per category. We generate these images using a single rigged model per animal species. For cow, sheep, and goat, we use a separate model for each sex category, incorporating major sex-specific attributes such as horns. Each model includes up to three different textures and 50 animated actions, such as running, walking, and drinking. We also randomly sample from a pool of 742 HDRI environmental maps to provide diverse lighting conditions for the training images. We then randomly sample camera viewpoints and poses from the animated actions to generate the training images. **??** showcases the horse model and some of the generated images used for training.

References

[1] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

[2] Tomas Jakab, Ruining Li, Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Farm3D: Learning articulated 3D animals by distilling 2D diffusion. In *Proc. 3DV*, 2024.

[3] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Proc. ICLR*, 2015.

[4] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proc. CVPR*, 2023.

[5] Nilesh Kulkarni, Abhinav Gupta, David F. Fouhey, and Shubham Tulsiani. Articulation-aware canonical surface mapping. In *Proc. CVPR*, pages 449–458, 2020.

[6] Zizhang Li, Dor Litvak, Ruining Li, Yunzhi Zhang, Tomas Jakab, Christian Rupprecht, Shangzhe Wu, Andrea Vedaldi, and Jiajun Wu. Learning the 3D fauna of the Web. In *Proc. CVPR*, 2024.

[7] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, 2022.

[9] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *Proc. ICLR*, 2021.

[10] Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Magicpony: Learning articulated 3d animals in the wild. *arXiv.cs*, 2022.

[11] Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht,

and Andrea Vedaldi. MagicPony: Learning articulated 3D animals in the wild. In *Proc. CVPR*, 2023.

[12] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements DINO for zero-shot semantic correspondence. *arXiv.cs*, abs/2305.15347, 2023.