SyDog: A Synthetic Dog Dataset for Improved 2D Pose Estimation

Moira Shooter Charles Malleson Adrian Hilton
Center of Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK
{m.shooter, charles.malleson, a.hilton}@surrey.ac.uk

UNIVERSITY OF SURREY
CVSSP Centre for Vision, Speech and Signal Processing
CVPR VIRTUAL JUNE 19-25
CV4Animals Workshop

1. Introduction

Motivation: Estimating the pose of animals from video helps to understand the animal motion and this supports many applications and disciplines such as:

- Veterinary sciences
- Biomechanical applications
- Neuroscience
- Robotics
- The entertainment industry

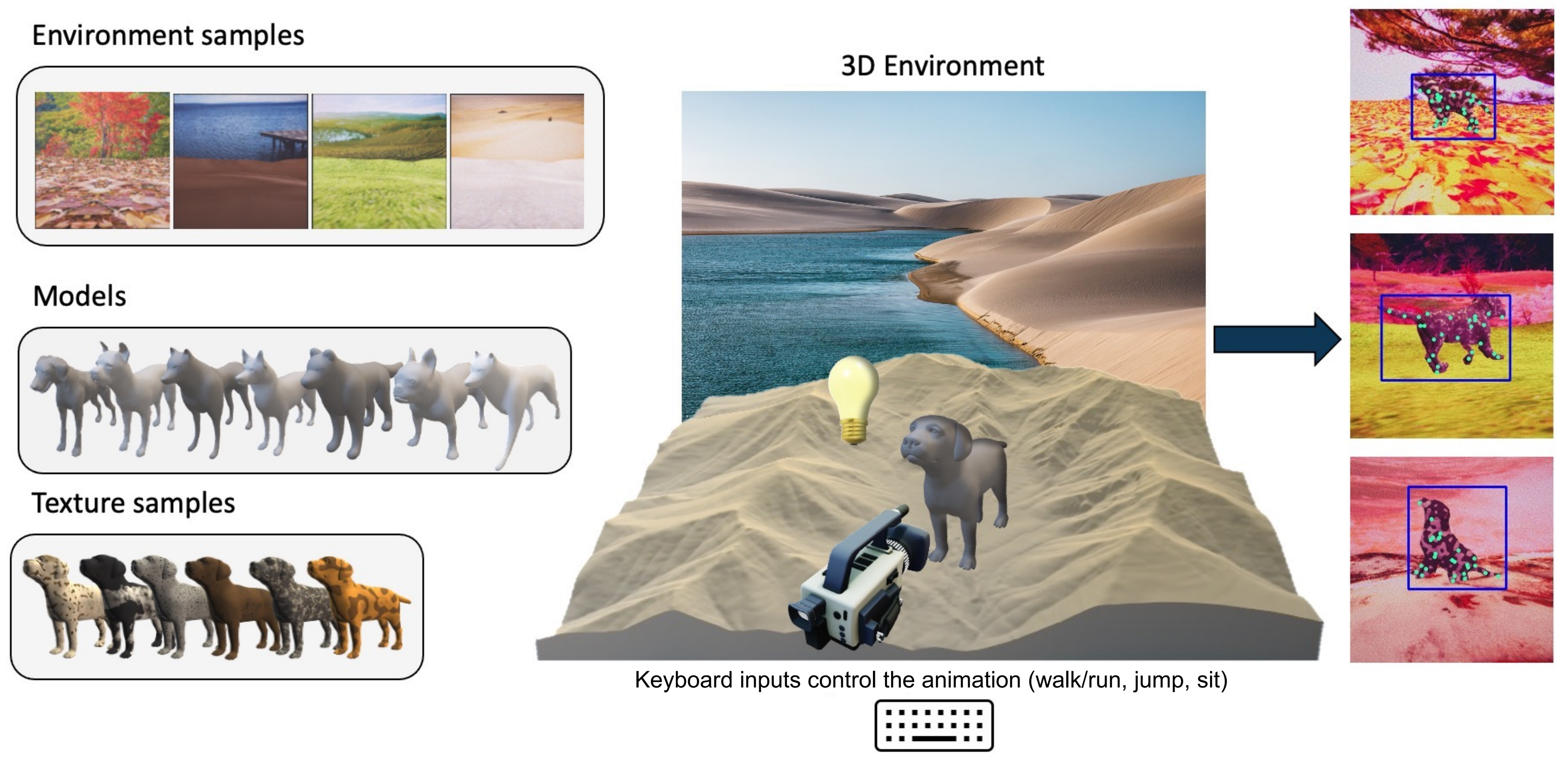
Problem: There is a lack of animal pose datasets. Manually annotating several keypoints in thousands of images is both labour intensive and expensive.

Question: Can synthetic data improve 2D pose estimation models?

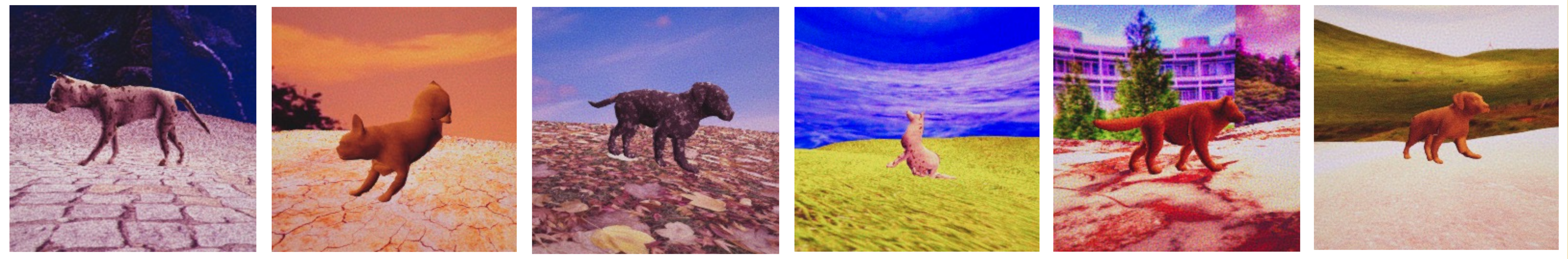
2. Data generation

- Generated image data of dogs in the game engine, Unity3D based on [Zhang et al. 2018]
- 8 dogs (3D models), 12 fur textures, 1k background images, 10 terrain textures, randomized light and camera settings
- **Output:** Images + 2D joint locations and bounding box coordinates

Input | Output

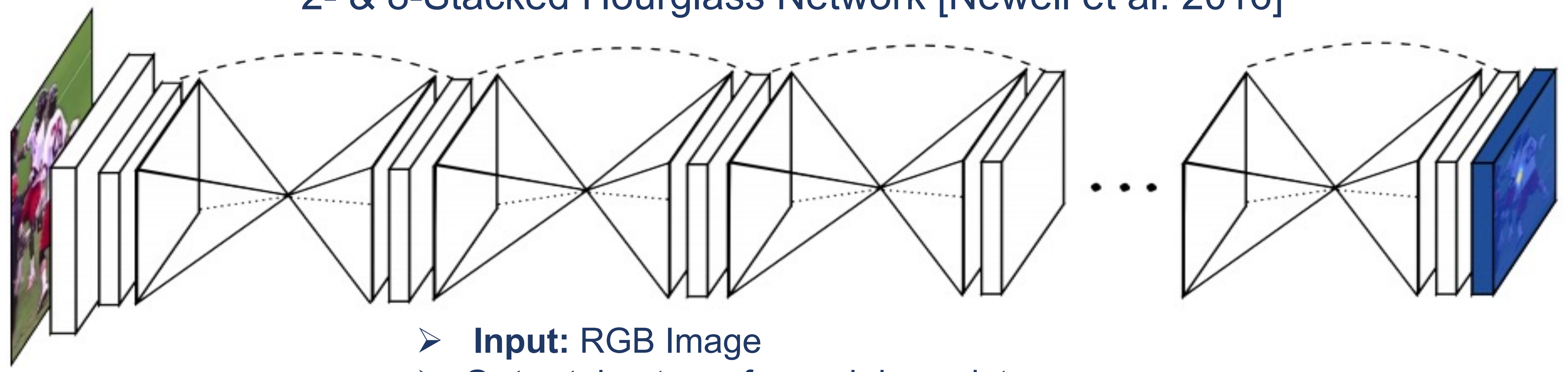


Samples from SyDog dataset



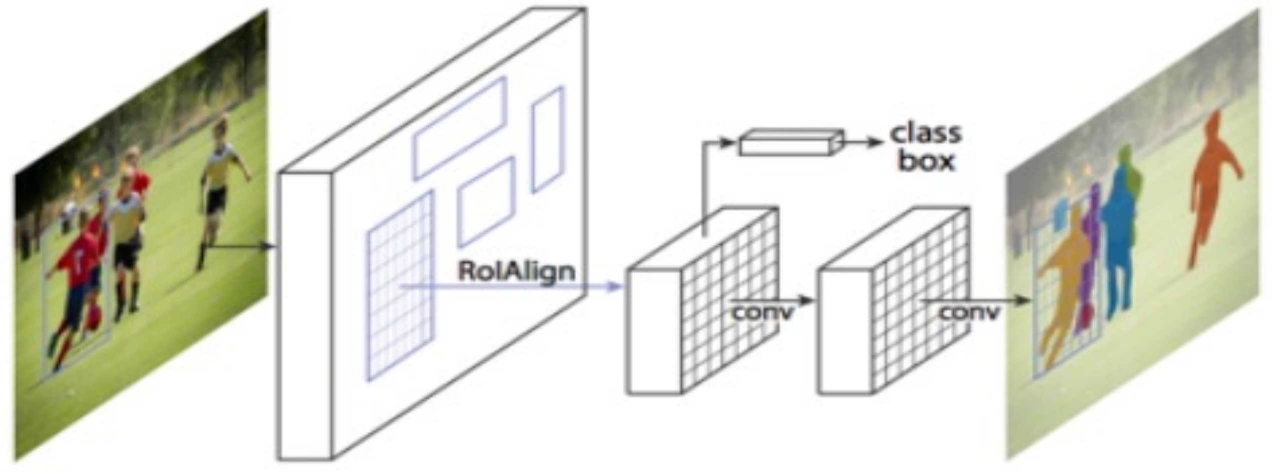
3. Experiments

2- & 8-Stacked Hourglass Network [Newell et al. 2016]



- **Input:** RGB Image
- **Output**: heatmap for each keypoint

Mask R-CNN [He et al. 2017]



- **Input:** RGB Image
- **Output**: keypoint coordinates instead of segmentation masks

Bridging the domain gap (synthetic vs real)

- **Fine-tuning** the networks trained with pure synthetic data with real data
- Training the networks with **a mixed dataset** (synth+real)

Evaluation Metrics

- Percentage of Correct Keypoints (PCK), at 10% of bounding box diagonal
- Mean Per Joint Position Error (MPJPE) w.r.t. bounding box diagonal

4. Results on SyDog [Ours]

Network	PCK (%)	MPJPE (%)
2HG	77.76	6.51
8HG	77.57	6.56
Mask R-CNN	68.98	11.02

Table 1: PCK@0.1 and MPJPE results from the 2-and 8-stacked hourglass network (2HG, 8HG) and the Mask R-CNN on the SyDog dataset

4. Results on StanfordExtra [Biggs et al. 2020]

Network	Dataset	Learning rate	PCK (%) ↑	MPJPE (%) ↓
2HG	Real	0.001	68.61	15.84
	Synthetic	0.001	16.20	46.26
	FT	0.001 → 0.001	76.57	11.80
	FT	0.001 → 0.000001	77.19	11.32
	Mixed@0.1	0.001	63.14	19.08
	Mixed@0.5	0.001	68.43	15.50
	Mixed@1.0	0.001	70.46	14.76
8HG	Real	0.001	68.90	15.64
	Synthetic	0.001	17.34	45.08
	FT	0.001 → 0.001	78.31	11.47
	FT	0.001 → 0.00001	78.65	11.19
	Mixed@0.1	0.001	65.04	17.81
	Mixed@0.5	0.001	71.76	15.19
	Mixed@1.0	0.001	72.09	14.97
Mask R-CNN	Real	0.00001	43.60	21.58
	Synthetic	0.001	13.22	37.49
	FT	0.00001 → 0.00001	50.77	20.03
	FT	0.00001 → 0.000001	46.58	21.17
	Mixed@0.1	0.001	41.27	22.82
	Mixed@0.5	0.001	47.71	21.64
	Mixed@1.0	0.001	45.77	21.61

Table 2: PCK@0.1 and MPJPE results from the 2-and 8-stacked hourglass network (2HG, 8HG) and the Mask R-CNN trained solely on (Real) and solely on the SyDog dataset (Synthetic) together with the fine-tuned (FT) models and the models trained with a mixed dataset (Mixed@fraction).

5. Conclusions

- We presented **a real-time system that generated 2D annotated images** containing dogs
- **We release SyDog**, a large-scale dataset of dogs with 2D keypoints and bounding box coordinates
- We demonstrated that using **the SyDog dataset improves the accuracy of pose estimation models** and reduces the need for labour intensive labelling

References

- He Zhang, Sebastian Starke, Taku Komura, and Jun Saito. Mode-adaptive neural networks for quadruped motion control. ACM Trans. Graph., 37(4), July 2018.
- Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. CoRR, abs/1603.06937, 2016.
- Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. CoRR, abs/1703.06870, 2017.
- Benjamin Biggs, Oliver Boyne, James Charles, Andrew Fitzgibbon, and Roberto Cipolla. Who left the dogs out?: 3D animal reconstruction with expectation maximization in the loop. In ECCV, 2020.