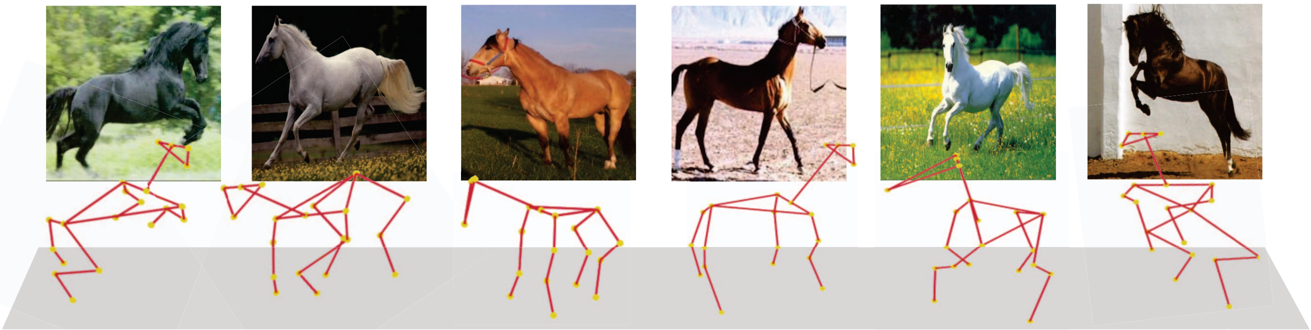
A Horse with no Labels: Self-Supervised Horse Pose Estimation from Unlabelled Images and Synthetic Prior

Jose Sosa, David Hogg
School of Computing, University of Leeds
{scjasm, D.C.Hogg}@leeds.ac.uk



Abstract

Obtaining labelled data to train deep learning methods for estimating animal pose is challenging. Recently, synthetic data has been widely used for pose estimation tasks, but most methods still rely on supervised learning paradigms utilising synthetic images and labels. Can training be fully unsupervised? Is a tiny synthetic dataset sufficient? What are the minimum assumptions that we could make for estimating animal pose? Our proposal addresses these questions through a simple yet effective self-supervised method that only assumes the availability of unlabelled images and a small set of synthetic 2D poses. We completely remove the need for any 3D or 2D pose annotations (or complex 3D animal models), and surprisingly our approach can still learn accurate 3D and 2D poses simultaneously. We train our method with unlabelled images of horses mainly collected for YouTube videos and a prior consisting of 2D synthetic poses. The latter is three times smaller than the number of images needed for training. We test our method on a challenging set of horse images and evaluate the predicted 3D and 2D poses. We demonstrate that it is possible to learn accurate animal poses even with as few assumptions as unlabelled images and a small set of 2D poses generated from synthetic data. Given the minimum requirements and the abundance of unlabelled data, our method could be easily deployed to different animals.

1. Introduction

One of the main bottlenecks for supervised animal pose estimation is obtaining pose annotations for training deep-learning models. While plenty of labelled data is available in the human domain, annotated animal datasets are scarce. In order to overcome the annotation issue, new alternatives have been adopted, such as training models with synthetic data.

We adapt a method from the human domain that learns human 3D poses from unlabelled images and a prior on 2D poses [27]. Our implementation translates this method to the animal domain, demonstrating that it applies to different body structures. Another essential addition to our approach is the origin of the 2D poses composing the prior. Unlike the original implementation [27], which uses a set of unpaired 2D poses from the training datasets, we further reduce the assumptions by using 2D poses from an existing CAD model of a horse [22]. Our model is unique in its simplicity compared with previous approaches for animal pose estimation with synthetic data. It does not require annotated training data. It uses only unlabelled images and a small set of synthetically generated 2D poses, which means that no synthetic images, pre-trained models, or complicated 3D models are required.

We train and test the model with unlabelled images of horses. Additionally, we use a prior on 2D pose generated from synthetic data [22]. By evaluating 2D and 3D predictions from our model, we demonstrate that our approach produces accurate 2D and 3D pose representations of the horses, although we are not using any annotations for the input images. Since the requirements for training our model are minimal, it could be easily applied to animal species with different body structures.

2. Related work

2.1. Animal pose estimation

Supervised deep learning methods for human pose estimation have been widely explored and perform well under different conditions [23, 29, 18]. However, in animal pose estimation, getting the labels needed for supervision is difficult in most cases. In particular, labelling key points is more expensive and laborious than producing other annotations, e.g. bounding boxes. On top of this, it would be infeasible to generate labelled data for the entire diversity of animal species in the world.

Since the 3D pose annotations are even more challenging to acquire than the 2D ones, many works on animal pose estimation have been focused only on estimating 2D pose [19, 20, 24, 26]. Not surprisingly, the backbones for most of these approaches are network architectures initially designed for the human domain, for example, stacked hourglass networks [23], ResNet [12], and OpenPose[6].

Although the problem of 3D animal pose estimation is more constrained and challenging, relevant work has also been carried out [1, 15, 11]. In this context, methods commonly rely on lower supervision levels to overcome the scarcity of labelled training data. For instance, the self-supervised approach of [33] estimates 3D pose for monkeys and dogs relying on multi-view supervision and a tiny portion of pose annotations. Dai *et al*. [8] proposes an similar method, but instead of multiview images, they assume the availability of actual 2D poses for each input image and lift these to 3D through self-supervision based on geometric consistency. Similar to [8] our method also estimates 3D pose using self-supervision with the same geometric consistency constraint. However, we learn the 2D and 3D poses directly from images in an end-to-end manner. Most importantly we do not require any annotations for the inputs.

2.2. Animal pose estimation with synthetic data

Synthetic data is a low-cost alternative to generate data with ground truth annotations with minimum effort. Recently, works on human [17, 10, 30] and animal pose [4, 22, 16, 6, 3, 2, 25, 36, 35, 28] estimation have adopted synthetic data to overcome the scarcity of keypoint labels.

Many animal pose estimation methods with synthetic data follow a supervised approach, meaning they use synthetically generated images and pose annotations for training. However, there is often a gap between synthetic and real data, so these approaches typically perform domain adaptation with samples from actual data. For example, [22] learns 2D pose for animals using images and labels generated from CAD models. They also incorporate a consistency-constrained semi-supervised method to adapt the predictions to real data. Similarly, [16] focuses on domain adaptation by generating pseudo-labels from the synthetic domain and then updating these to match the actual data. Unlike these approaches, our formulation helps to reduce the complexity and requirements for training even more. It is as simple as using unlabelled real images and a set of synthetically generated 2D poses, i.e. there is no need to generate pictures from the synthetic data. Furthermore, an adversarial loss helps to learn poses that do not necessarily appear in the prior of synthetic 2D poses without having additional processes to align domains.

More related to our work, [28] relies on a self-supervised method that assumes synthetic 2D poses and real images for estimating 2D mouse pose. However, we advance [28] by incorporating geometry consistency, allowing our model to further estimate 3D pose.

Synthetic data also plays an essential role in several works that learn richer structures, such as animal shapes, mainly for different quadrupeds like dogs [2, 25, 3], tigers, lions, horses [36], and zebras [35]. However, the success of these approaches is constrained by having access to sophisticated and expensive animal models, which is not required in our approach.

3. Method

The method is essentially that from [27]. Unlike Sosa [27] we translate this method to the animal domain and most importantly we change the origin of the prior. In the original paper they use a prior of 2D poses coming from unpaired annotations of the training dataset. We remove this by generating the 2D poses from synthetic data. We reproduce the method here so that the current paper is self-contained.

The main component of the approach is an image to 3D pose mapping, indicated with a dotted box in Figure 1. The first part of this mapping employs a CNN Φ to map the input image x to an intermediate skeleton image s. Then, another CNN Ω maps s to a 2D pose representation y. In the final stage, y is mapped to the 3D pose v by means of a fully connected network Λ. For training this set of networks, we incorporate it within a larger structure which allows for self supervision. In particular, we rely in a loop of transformations of the 3D pose v. We also use a discriminator D together with the prior on synthetic 2D poses, to ensure that the generated skeletons s are realistic.

3.1. Main mapping

The image to pose mapping consists of 3 networks Φ, Ω, and Λ that allows the input image x to be mapped to its 3D pose representation v. This mapping also produces two

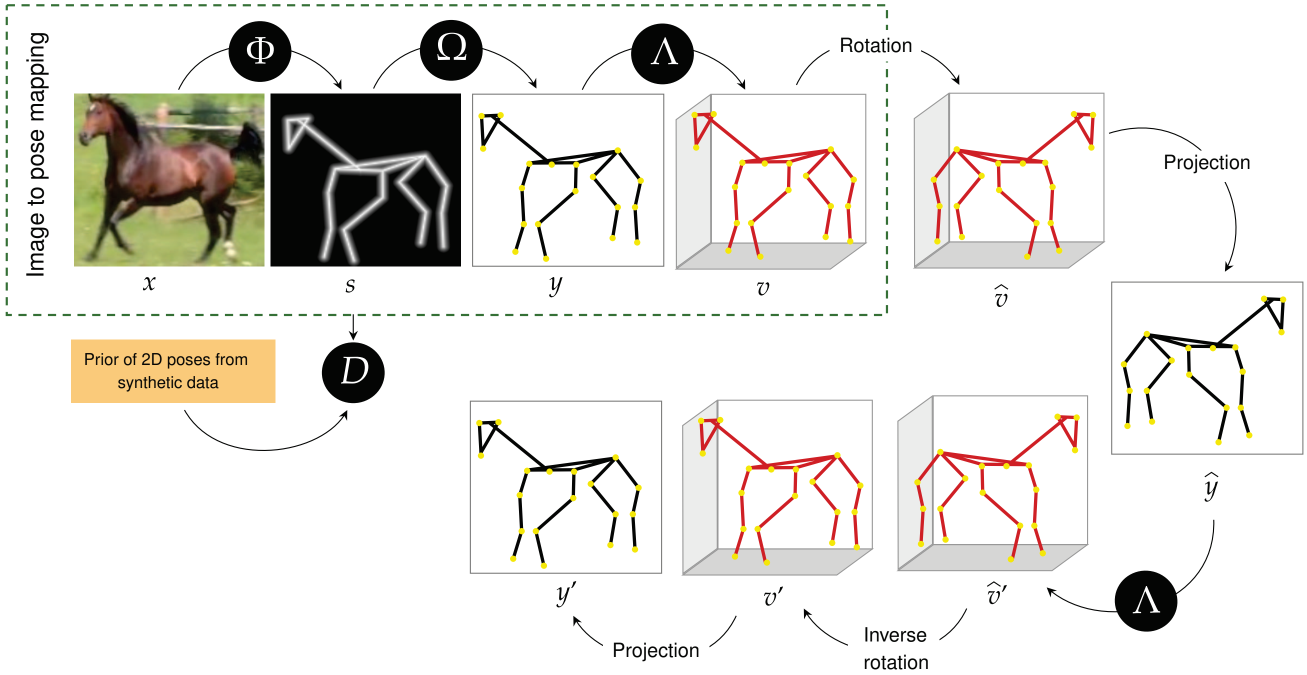


Figure 1. The method is an adaptation of [27]. The main difference is the origin of the prior on 2D poses. In this case we use a publicity available set of 2D poses from an existing CAD model of a horse from [22].

intermediate representations of the input, a skeleton image s, and a 2D pose y. Specifically, Φ learns to align the input image with its respective skeleton image representation, i.e. $s = \Phi(x)$. Then, Ω learns to extract keypoints from s, obtaining a 2D pose as output $y = \Omega(\Phi(x))$. Finally, Λ acts as a lifter of the 2D pose y to get the 3D pose v. For each pair of joint positions (x_i, y_i) in y, the network estimates a depth $z_i = d + \Delta$, where Δ is a constant depth.

Overall, we use the same network structure as in [27] with exception of Λ. Since we are not trying to learn elevation angles for the geometry transformations like [27, 31], we opt for a simpler structure as in [7, 18].

3.2. Self-supervision

As illustrated by Figure 1, we include the main mapping within a large network structure that allows to self-supervise the training. This structure uses a discriminator network D, which relies on a prior of synthetic 2D poses to help the mapping produce skeleton images that are as realistic as possible. Furthermore, it incorporates a loop of random rotations and projections of the 3D pose v to ensure geometry consistency for the 3D predictions.

3.2.1 Synthetic pose prior

To create the prior of 2D poses, we use a publicly available dataset of synthetic 2D poses generated from a CAD model of a horse [22]. The prior is needed during training to ensure the estimated skeleton image looks as realistic as possible. Note that generating the prior from synthetic data and not from annotations of the dataset like [27] provides more flexibility to the method to be trained with completely unlabelled datasets, which are abundant in the animal domain. Our synthetic prior contains around 10k different 2D poses, representing approximately one-third of the available images for training. Figure 2 provides examples of some 2D poses p in the prior.

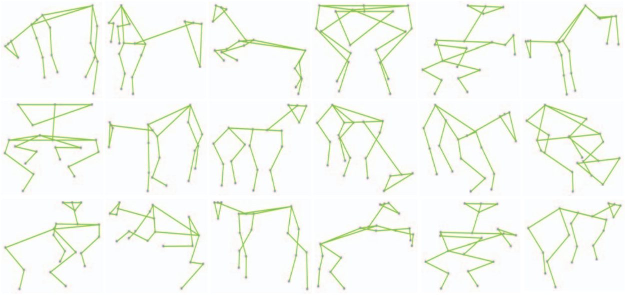


Figure 2. Random 2D poses from the syntethic prior.

The purpose of having a prior of 2D poses is to use these as a reference distribution for the discriminator network D. Since our implementation of D works directly with images, we must first render the synthetic 2D poses to skeleton images. This is done by using the rendering function β from [14], which given a set of 2D joint positions p and their connections, can generate a skeleton image $w = \beta(p)$. Then, the goal of D is to evaluate whether or not the predicted skeleton image $s = \Phi(x)$, looks like an authentic skeleton image w such as those in the prior. Following [27, 14] we use an adversarial loss to compare w and s:

$$L_D = \mathbb{E}_w(log(D(w)) + \mathbb{E}_s(log(1 - D(s)) \tag{1}$$

3.2.2 Geometry consistency

We rely on the idea of geometric consistency from [7] to facilitate the learning of the lifting network Λ and, therefore, the whole mapping. Essentially this involves a series of rotations and projections of the 3D pose v. First, v is randomly rotated to $\hat{v}$ using a rotation matrix, which is constructed by sampling azimuth and elevation angles from a fixed uniform distribution [7]. Then, $\hat{v}$ is projected to a 2D pose $\hat{y}$. Given the projection of the rotated 3D pose $\hat{v}$, the same lifting network Λ estimates its 3D representation $\hat{v}'$. Lastly, the inverse rotation is applied to the 3D pose $\hat{v}'$ to obtain v', and v' is projected to 2D to get the 2D pose y'.

After the loop of projections and rotations we expect the poses on the forward and backward parts to be as similar as possible. For example, the 3D poses v and v' should be similar, and the same with $\hat{v}$ and $\hat{v}'$. This also applies to the 2D poses y and y'. Therefore, we can derive the following component loss functions:

$$L_{2D} = ||y' - y||^2 \tag{2}$$

$$L_{3D} = ||(v'^{(j)} - v'^{(k)}) - (v^{(j)} - v^{(k)})||^2 \tag{3}$$

$$L_{r3D} = ||\hat{v}' - \hat{v}||^2 \tag{4}$$

Note that for Equation 3 we follow [31, 27] and instead of comparing the v and v' with a L_2 loss we measure the degree of deformation between 3D poses using two samples j and k in a batch. For simplicity, we refer to the sum of these three losses as L_{GC} given by

$$L_{GC} = L_{2D} + L_{3D} + L_{r3D} \tag{5}$$

3.2.3 Training and additional losses

Following [14] we include an extra loss term L_Ω to evaluate the mapping $y = \Omega(s)$, i.e. from the skeleton image s to the 2D pose y.

$$L_\Omega = ||(\Omega(\beta(p)) - p)||^2 + \lambda||\beta(y) - s||^2 \tag{6}$$

where λ represents a balancing coefficient, p is a 2D pose from the unpaired prior and β is rendered from [14].

We train all the networks from scratch using a loss function L consisting of three components from Equation 1, Equation 5, and Equation 6.

$$L = L_D + L_{GC} + L_\Omega \tag{7}$$

At inference time, we only keep the elements from the main mapping as illustrated in the dotted box from Figure 1, i.e. the loop of rotations and projections, and D are only needed during training.

4. Experiments

4.1. Data

We train the model with a dataset of video frames depicting full-body horses. First, we select the horse subset from the latest version of the TigDog dataset [9]. We use video frames for all the horse sequences in the dataset, discarding video frames showing partially visible horses. To increase the diversity of horses in the training set, we automatically collect video frames for a manually defined group of YouTube videos that are expected to show horses throughout (i.e, in most video frames). To gather the video frames automatically from a video, we follow a three-step process:

1. Download the video from YouTube and split it into frames.

2. Process each frame using a pre-trained model from [32], which identifies the horse and produces a segmentation mask. Remove frames that do not contain a horse.

3. Resize the frames showing a horse to a predefined size (128 × 128) and save them along with their respective segmentation mask generated by the model.

We collect frames containing complete horses from about 60 videos, representing 47k frames (plus around 6k from [9]). Note that this dataset is relatively small compared to what is required for training human pose estimation models — our horse dataset is only 1.3% of the size of the Human3.6M dataset [13] and 3.6% of the size of the MPI-INF-3DHP dataset [21].

4.1.1 Test data

In real life applications we cannot assume that test data will come from the same source as the training data. Thus, instead of selecting a hold-out set of frames for each training video (which potentially could lead to better performance), we use more challenging data to test our model. In particular, we utilise images from a different collection: the Weizmann dataset [5]. However, this data does not contain annotations for 2D or 3D poses. We therefore manually annotate the 2D poses consisting of 15 joint positions (3 for each front and rear limb, 1 for the chin, and 2 for the eyes) for all the images in the Weizmann dataset showing

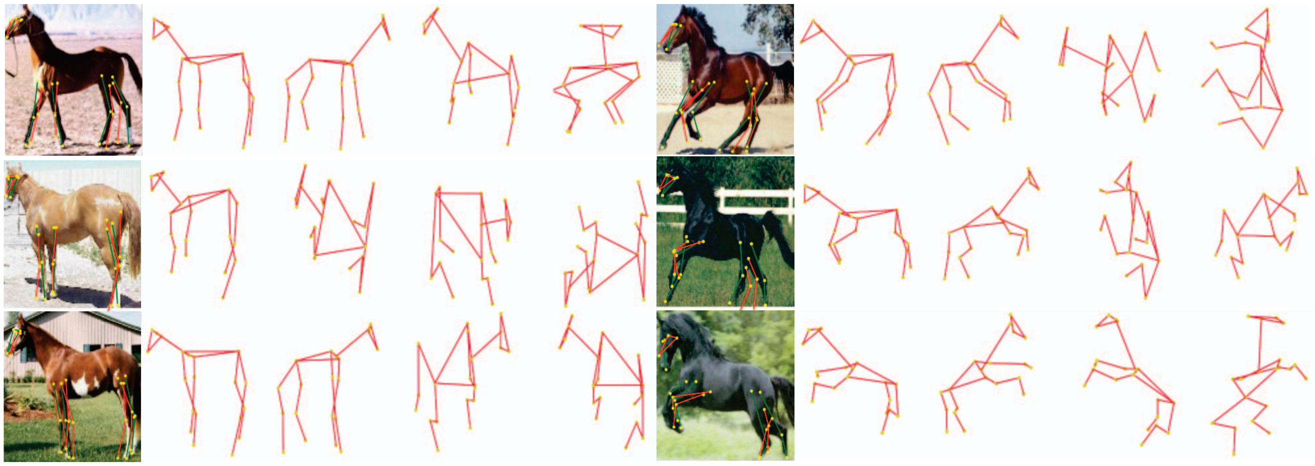


Figure 3. **3D poses estimated by our method.** The first and sixth columns show the images with their estimated (red) and ground truth (green) 2D poses. The rest of the columns illustrate novel views of the predicted 3D pose.

full-body horses (around 300). We use the pose annotations as ground truth to quantitatively evaluate the estimated 2D poses from our method.

4.2. Evaluation

Since there is a lack of available horse datasets with 3D pose annotations for a quantitative performance evaluation, we only assess the quality of the 3D predictions qualitatively. While obtaining 2D poses is more feasible than 3D poses, we evaluate the emergent 2D pose predictions y quantitatively and qualitatively. Note that although the goal of the model is predicting 3D poses, the emergent 2D pose representations are also worth evaluating. We assume that if the 2D poses are good, it is very likely the 3D poses will be reasonable as well.

In line with previous works for 2D animal pose estimation, we use the Percentage of Correct Keypoints (PCK@0.05) to quantitatively evaluate our 2D predictions. Our predicted poses are composed of 20 joint positions. However, we use only 15 in order to compare with the ground truth 2D poses from the Weizmann dataset.

4.3. Results & Discussion

4.3.1 Results on 3D pose predictions

Given the scarcity of ground truth 3D data for horses, we provide only a qualitative evaluation of the 3D poses estimated by our trained model in Figure 3.

Additionally, we test the generalisation capability of our model by evaluating it on a dataset of zebras [35]. Because of the anatomical similarities between zebras and horses, the trained model with the horse data can still estimate plausible 3D poses for zebras (although it has never seen a zebra during training). Figure 4 displays some 3D predictions for zebras. Given the slight differences between the two species (zebras having slightly wider chests and shorter legs), these results show the robustness of our method to different domains.

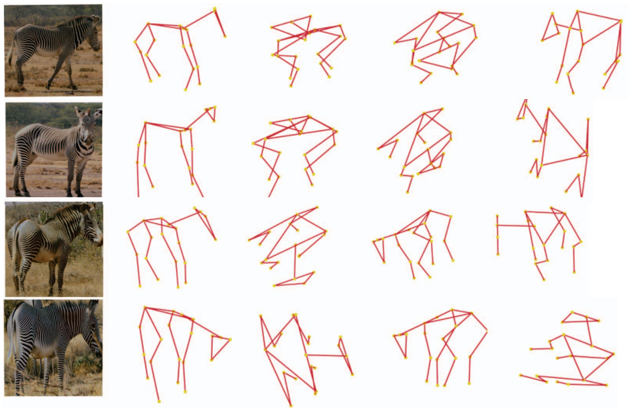


Figure 4. **3D pose predictions for zebras.** First column shows the input image. Following columns show novel views of the 3D poses.

4.3.2 Results on 2D pose predictions

Using the trained model, we produce 2D poses for all the images in the test set. Each pose prediction consists of 20 joint positions. However, when comparing against ground truth, we only keep 15 joint positions to match the annotations. Figure 5 shows some predicted 2D poses by our model compared with their respective ground truth.

In addition, we reproduce the method from [28] that originally estimates 2D poses for mice. We train it with the same assumptions that our method, i.e. our same horse dataset and synthetic 2D poses. We use the Weizmann dataset to evaluate and compare their predictions with the ones obtained with our 3D method. As illustrated by Fig-

Method	Evaluation Data	Eyes	Chin	Shoulders	Knees	Hooves	**Mean**
Syn - Mu *et al.* [22]	TigDog dataset	46.08	53.86	20.46	24.20	17.45	25.33
Sosa [28]	Weizmann dataset	45.67	44.67	33.00	37.67	26.67	37.54
CycleGAN [34]	TigDog dataset	70.73	84.46	56.97	49.91	35.95	51.86
Ours	Weizmann dataset	49.3	58.3	34.2	44.7	31.2	43.50

Table 1. **Horse 2D pose estimation accuracy.** We calculate the accuracy of our predicted 2D poses using the PCK@0.05 metric. For each image in the Weizmann dataset, the predicted 2D pose is compare against its respective ground truth. We also list some works that estimate 2D poses using synthetic data.

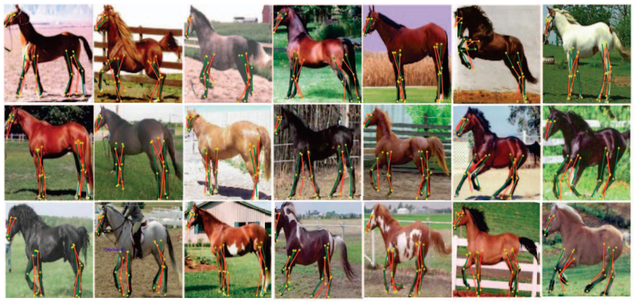


Figure 5. **Predicted 2D poses with our method.** Each image comes from the test set; the red lines represent the connections between our method's estimated 2D joint positions. The green lines represent the connections between the ground truth joint positions.

ure 6, our model for 3D poses can produce more accurate 2D pose representations than the 2D pose estimator from [28]. This comparison demonstrates the value of incorporating the geometry consistency idea for lifting 2D poses to 3D.

We use the PCK@0.05 metric to evaluate the predicted 2D poses against their respective ground truth. Table 1 shows the accuracy results of our quantitative evaluation for 2D pose. It also includes results for approaches that work under similar conditions. However, note that except for [28], which assumes the same setting as our method, the others methods apply supervised learning during training. Although the performance is not better than some of the methods listed in the table, it is also competitive, given the minor requirements of our method.

Furthermore, we experiment by training our method on the synthetically generated images of zebras from [35], and using the same synthetic 2D horse poses as prior. We then test on the same dataset of real zebras [35] as in previous experiments (model trained with horse images and synthetic 2D poses as prior). Despite the differences between domains, the model trained with purely synthetic data (synthetic images of zebras and synthetic 2D poses of horses) produces similar 2D poses as the model trained with real horse images and the synthetic 2D horse poses. Figure 7 shows the predicted 2D poses for different images from both configurations.

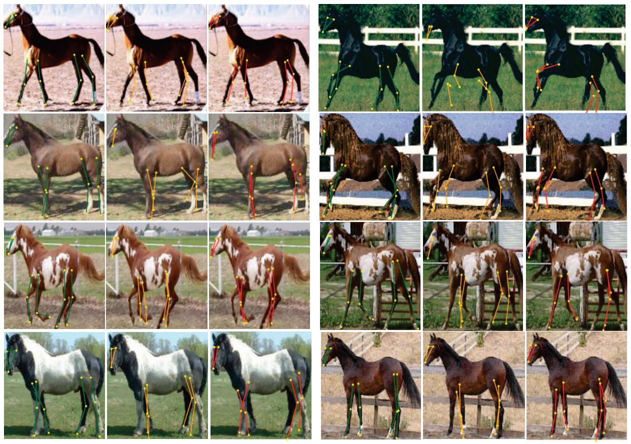


Figure 6. **Comparison of 2D pose predictions with a similar method.** The first and fourth columns show the ground truth joint positions (green). The second and fifth columns show the estimated 2D poses by [28] (orange). The third and sixth columns display the estimated 2D poses by our method (red).

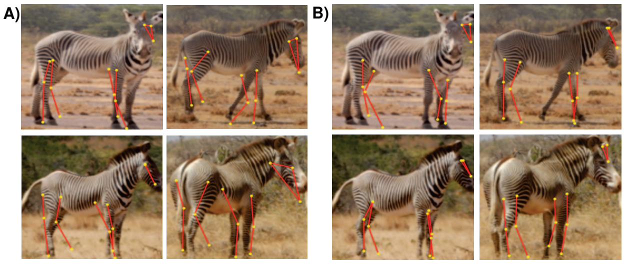


Figure 7. **Predicted 2D poses.** Block A, shows 2D poses predicted by the model trained with images of horses and the prior on synthetic 2D poses. Block B, display 2D predictions using synthetic images of zebras and a the same prior on synthetic 2D poses from horses.

4.3.3 Failed cases

Note that the quality of the emergent 2D pose estimations influences the accuracy of the final 3D pose predictions. Therefore, when the previous 2D predictions depict proper horse poses, the 3D predictions are expected to be more accurate. Surprisingly, even for some non-accurate 2D predictions, our model can still recover a plausible 3D horse pose, as shown in Figure 8.

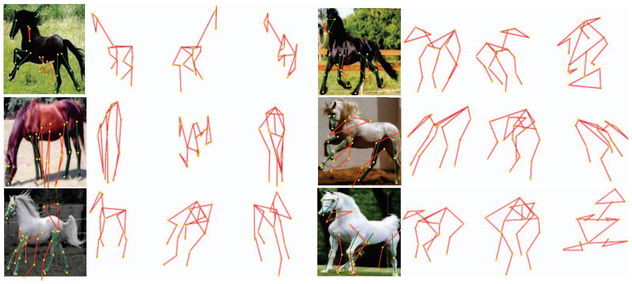


Figure 8. **Failed cases for the estimated 2D poses.** We select the 2D poses with lower accuracy (PCK@0.05) to have a look at their respective 3D predictions.

5. Conclusion

We have successfully adapted a method originally designed to estimate 3D human poses to the animal domain. We further reduce its requirements by generating the needed prior from synthetic data. We show that with only unlabelled images and a small set of synthetic 2D poses, it is possible to learn 3D representations. By reducing the data requirements for training to a minimum, our proposal could be applied to many unlabelled detests without collecting annotations needed for supervised training.

From our results, there is clearly room for further improvement. Two ideas for exploration in future are (1) to incorporate temporal information into the approach, and (2) to follow previous work in fine-tuning with small amounts of actual data to reduce the gap between the synthetic and real domains.

Acknowledgments

Thanks to Rebecca Stone from the School of Computing at University of Leeds for useful feedback and great discussions.

References

[1] Praneet C Bala, Benjamin R Eisenreich, Seng Bum Michael Yoo, Benjamin Y Hayden, Hyun Soo Park, and Jan Zimmermann. Automated markerless pose estimation in freely moving macaques with openmonkeystudio. *Nature communications*, 11(1):4560, 2020. 2

[2] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 195–211. Springer, 2020. 2

[3] Benjamin Biggs, Thomas Roddick, Andrew Fitzgibbon, and Roberto Cipolla. Creatures great and smal: Recovering the shape and motion of animals from video. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part V 14*, pages 3–19. Springer, 2019. 2

[4] Luis A Bolaños, Dongsheng Xiao, Nancy L Ford, Jeff M LeDue, Pankaj K Gupta, Carlos Doebeli, Hao Hu, Helge Rhodin, and Timothy H Murphy. A three-dimensional virtual mouse generates synthetic training data for behavioral analysis. *Nature methods*, 18(4):378–381, 2021. 2

[5] Eran Borenstein, Eitan Sharon, and Shimon Ullman. Combining top-down and bottom-up segmentation. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pages 46–46. IEEE, 2004. 4

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 2

[7] Ching-Hang Chen, Ambrish Tyagi, Amit Agrawal, Dylan Drover, Rohith Mv, Stefan Stojanov, and James M Rehg. Unsupervised 3d pose estimation with geometric self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5714–5724, 2019. 3, 4

[8] Xiaowei Dai, Shuiwang Li, Qijun Zhao, and Hongyu Yang. Unsupervised 3d animal canonical pose estimation with geometric self-supervision. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2023. 2

[9] Luca Del Pero, Susanna Ricco, Rahul Sukthankar, and Vittorio Ferrari. Behavior discovery and alignment of articulated object classes from unstructured video. *International Journal of Computer Vision*, 121:303–325, 2017. 4

[10] Carl Doersch and Andrew Zisserman. Sim2real transfer learning for 3d human pose estimation: motion to the rescue. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[11] Adam Gosztolai, Semih Günel, Victor Lobato-Ríos, Marco Pietro Abrate, Daniel Morales, Helge Rhodin, Pascal Fua, and Pavan Ramdya. Liftpose3d, a deep learning-based approach for transforming two-dimensional to three-dimensional poses in laboratory animals. *Nature methods*, 18(8):975–981, 2021. 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[13] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 4

[14] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Self-supervised learning of interpretable keypoints from unlabelled videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8787–8797, 2020. 3, 4

[15] Daniel Joska, Liam Clark, Naoya Muramatsu, Ricardo Jericevich, Fred Nicolls, Alexander Mathis, Mackenzie W. Mathis, and Amir Patel. Acinoset: A 3d pose estimation dataset and baseline models for cheetahs in the wild. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13901–13908, 2021. 2

[16] Chen Li and Gim Hee Lee. From synthetic to real: Unsupervised domain adaptation for animal pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1482–1491, 2021. 2

[17] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 2

[18] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2640–2649, 2017. 2, 3

[19] Alexander Mathis, Thomas Biasi, Steffen Schneider, Mert Yuksekgonul, Byron Rogers, Matthias Bethge, and Mackenzie W Mathis. Pretraining boosts out-of-domain robustness for pose estimation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1859–1868, 2021. 2

[20] Alexander Mathis, Pranav Mamidanna, Kevin M Cury, Taiga Abe, Venkatesh N Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience*, 21(9):1281–1289, 2018. 2

[21] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017. 4

[22] Jiteng Mu, Weichao Qiu, Gregory D Hager, and Alan L Yuille. Learning from synthetic animals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12386–12395, 2020. 1, 2, 3, 6

[23] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 483–499. Springer, 2016. 2

[24] Talmo D Pereira, Diego E Aldarondo, Lindsay Willmore, Mikhail Kislin, Samuel S-H Wang, Mala Murthy, and Joshua W Shaevitz. Fast animal pose estimation using deep neural networks. *Nature methods*, 16(1):117–125, 2019. 2

[25] Nadine Rüegg, Shashank Tripathi, Konrad Schindler, Michael J Black, and Silvia Zuffi. Bite: Beyond priors for improved three-d dog pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8867–8876, 2023. 2

[26] Helena Russello, Rik van der Tol, and Gert Kootstra. T-leap: Occlusion-robust pose estimation of walking cows using temporal information. *Computers and Electronics in Agriculture*, 192:106559, 2022. 2

[27] Jose Sosa and David Hogg. Self-supervised 3d human pose estimation from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4787–4796, 2023. 1, 2, 3, 4

[28] Jose Sosa, Sharn Perry, Jane Alty, and David Hogg. Of mice and pose: 2d mouse pose estimation from unlabelled data and synthetic prior. *arXiv preprint arXiv:2307.13361*, 2023. 2, 5, 6

[29] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014. 2

[30] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 109–117, 2017. 2

[31] Bastian Wandt, James J Little, and Helge Rhodin. Elepose: Unsupervised 3d human pose estimation by predicting camera elevation and learning normalizing flows on 2d poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6635–6645, 2022. 3, 4

[32] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019. 4

[33] Yilun Zhang and Hyun Soo Park. Multiview supervision by registration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 420–428, 2020. 2

[34] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 6

[35] Silvia Zuffi, Angjoo Kanazawa, Tanya Berger-Wolf, and Michael J Black. Three-d safari: Learning to estimate zebra pose, shape, and texture from images" in the wild". In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5359–5368, 2019. 2, 5, 6

[36] Silvia Zuffi, Angjoo Kanazawa, and Michael J Black. Lions and tigers and bears: Capturing non-rigid, 3d, articulated shape from images. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3955–3963, 2018. 2