DigiDogs: Single-View 3D Pose Estimation of Dogs using Synthetic Training Data

Moira Shooter
m.shooter@surrey.ac.uk

Charles Malleson
charles.malleson@surrey.ac.uk

Adrian Hilton
a.hilton@surrey.ac.uk

Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey,
Guildford UK

Abstract

We propose an approach to automatically extract the 3D pose of dogs from single-view RGB images using only synthetic data for training. Due to the lack of suitable 3D datasets, previous approaches have predominantly relied on 2D weakly supervised methods. While these approaches demonstrate promising results, some depth ambiguities still persist indicating the neural network's limited understanding of the 3D environment. To tackle these depth ambiguities, we generate a synthetic 3D pose dataset (DigiDogs) by modifying the popular video game Grand Theft Auto. Additionally, to address the domain gap between synthetic and real data, we harness the power of Meta's foundation model DINOv2 due to its generalisation capability and fine-tune it for the application of 3D pose estimation. Through a combination of qualitative and quantitative analyses, we demonstrate the viability of estimating the 3D pose of dogs from real-world images using synthetic training data.

1. Introduction

Automatically extracting the 3D pose of animals from images holds immense potential across various fields, including ecology, biology, and wildlife conservation. Scientists can use this technology to study animal behaviour and movement patterns in their natural habitats, gaining valuable insights into the ecosystem and supporting biodiversity preservation. Additionally, the digitisation of animals has a significant impact on the emerging field of metaverse development. Through precise and lifelike 3D animal reconstructions, a more authentic and immersive virtual experience can be achieved, enhancing the overall credibility and engagement within the virtual world.

The field of 3D human pose estimation has advanced rapidly due to the availability of 3D datasets. In contrast, the progress in 3D animal reconstruction has been comparatively slow due to the scarcity of datasets, this is primarily due to the inherent difficulty in capturing 3D datasets for animals. As a consequence, researchers have faced challenges in obtaining sufficient and diverse data to train accurate animal 3D pose estimation models. To tackle this issue, previous research has predominantly relied on 2D weak supervision methods [4, 12, 24]. These methods do not have inherent 3D knowledge, so incorporating prior knowledge such as the animal's shape is necessary to mitigate the ambiguity of single-view 2D-to-3D mapping and improve the accuracy of 3D pose estimations. While these approaches have demonstrated promising results, the 3D pose often remains incomplete as the neural network lacks a comprehensive understanding of the 3D environment. This limitation becomes apparent when viewing the reconstruction from angles which are different from the original 2D camera view direction.

To the best of our knowledge, the only publicly available 3D dog pose dataset is RGBD-Dog [14]. The dataset includes 7 different types of dogs wearing motion capture (mocap) suits. The data was captured using multiple Microsoft Kinect v2s along with a mocap system, providing accurate 3D ground truth skeletons. However, due to the controlled nature of the data collection, a model trained on this dataset may face challenges when reconstructing poses from in-the-wild RGB images, as the dataset lacks the necessary diversity in, for example, background and illumination. To increase the diversity in the dataset and to address the lack of 3D pose datasets, we propose to generate and use synthetic data for training similarly to [11, 20]. However, using synthetic training data alone may lead to poor inference performance on real data, due to the domain gap between real and synthetic images.

We address two key problems for the 3D pose estimation of dogs from images. Firstly, in order to tackle the problem of in-the-wild 3D datasets and enhance the robustness of 3D pose estimation, we present a novel 3D synthetic dataset (DigiDogs) that consists of a variety of dog videos accompanied by both 2D and 3D ground truth labels. It was generated by modifying the game Grand Theft Auto (GTA) [1]. The dataset features a diverse collection of videos showcasing 8 distinct dogs engaged in various activities. Each video

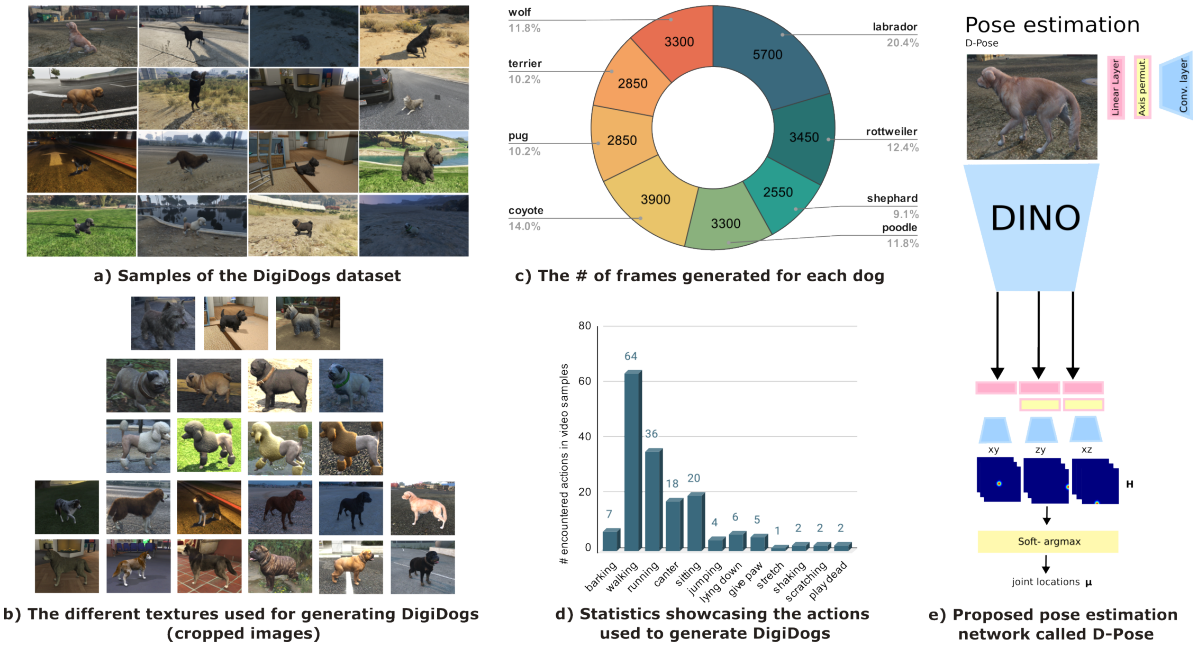

Figure 1. a) Illustrates samples of the DigiDogs dataset, b) showcases the different textures used for generating the DigiDogs dataset, c) is a pie chart showing the number of frames generated using the type of dog breed, d) shows the number of encountered actions across the videos generated and e) shows the proposed network used to estimate the 3D pose.

captures a single dog, ensuring individuality in breed, size, and texture representation. Moreover, the dataset offers a rich variety of backgrounds, encompassing both indoor and outdoor settings, along with realistic scene illuminations. Secondly, to address the domain gap we propose to build upon and fine-tune the foundation model, DINOv2 [22] directly on the pose estimation task. The **key contributions** of our work are:

- The generation and release of a synthetic 3D pose dataset called DigiDogs compromising of videos of dogs in various scenes from the game Grand Theft Auto.
- We address the domain gap between synthetic and real data by leveraging a foundation model (DINOv2).
- To the best of our knowledge, we are the first to estimate the 3D pose of dogs from in-the-wild monocular images using synthetic data alone.

2. Related work

2.1. Animal 3D pose estimation

3D pose estimation methods can be broadly categorised into two groups: multi-view methods and single-view methods. Although multi-view 3D pose estimation remains the most accurate approach for extracting the 3D pose of animals, it necessitates bringing the animals into a laboratory and capturing them from multiple synchronised views using motion capture technology [3, 19]. This method poses challenges related to setting up the mocap system and most likely dealing with the uncooperative and natural behaviour of animals. To overcome these difficulties, prior research in the field of 3D animal reconstruction has increasingly focused on single-view reconstruction techniques, primarily relying on 2D supervision as an alternative approach, considering that acquiring 2D datasets is easier than acquiring 3D datasets.

Relying solely on 2D ground truth can introduce depth ambiguity issues, necessitating the incorporation of prior knowledge to mitigate this problem. Previous methods have addressed this challenge by using the animal parametric model SMAL [34] as a geometric prior. Inspired by the human-specific parametric model SMPL [18], SMAL was learned from 41 scanned toy figures of various quadrupeds. Various versions of the SMAL model have since been released for different animal cases, including horses [16], birds [2, 30], and dogs [4, 7, 24, 26]. Additionally, more relevant to our research, Kearney *et al.* [14] introduced the RGBD-Dog dataset, the only publicly available 3D dataset featuring multi-view videos of dogs recorded in mocap suits. This dataset serves as a quantitative evaluation benchmark for future approaches and has been used to reconstruct dogs from depth images. Although it provides high-quality 3D ground truth, neural networks trained using the RGB images from the RGBD-Dog dataset may exhibit poor gen-

eralisation to in-the-wild scenarios, as these images feature dogs wearing mocap suits and were recorded indoors in a controlled environment. Unlike methods utilising the parametric model as a geometric prior, our approach stands out by not relying on any priors. Instead, we directly train our neural network using 3D pose data. Furthermore, our method incorporates 3D data from a diverse dataset instead of a mocap indoor dataset, leading to increased generalisation performance for in-the-wild scenarios.

2.2. Domain adaptation

The use of synthetic data as training data has become increasingly popular [9, 29, 32] due to its numerous benefits, such as the ability to generate an unlimited amount of data and having complete control over the data generation process. Although the setup for generating synthetic data can be time-consuming, it is often considered a preferable alternative to tedious and labour-intensive labeling work. In the realm of animal-related research synthetic data usage is not uncommon. In fact, [20] were pioneers in using CAD models to generate synthetic data for animal pose estimation. Similarly, [27] created a 2D pose dataset to address the pose estimation problem for images of dogs. Another recent [6] study employed image-domain translation techniques to produce realistic videos of mice for 2D/3D pose estimation model training, achieving comparable accuracies to models trained solely on real-world data and showcasing the potential of synthetic data in animal pose estimation. Biggs *et al*. [5] performed 3D reconstruction of various animals by registering meshes to synthetically generated silhouettes, while Zuffi *et al*. [33] successfully achieved the automatic reconstruction of 3D pose, shape, and appearance of in-the-wild images of zebras.

Recently, foundation models such as the Contrastive Language-Image Pre-training model (CLIP) [23] and DINOv2 [22] have gained significant popularity as foundation models across various visual tasks, thanks to their robust pretraining capabilities and impressive state-of-the-art performance. The following methods have benefited from these models for 3D hand pose estimation [15], generating animations from images [13] and generating 3D motions from text prompts [28]. In our work, we harness the benefits of the DINOv2 model [22] for the 3D pose estimation task to narrow the domain gap. To the best of our knowledge, we are the first to estimate the 3D pose of dogs from in-the-wild monocular images using only synthetic 3D training data.

3. DigiDogs dataset generation

The dataset was generated by modifying the game, Grand Theft Auto (GTA), which simulates a world based on Los Angeles. We decided to generate the data using GTA due to it being near photorealistic. Parameters such as the time, lighting, weather, camera views and location (indoor/outdoor) were randomised. The human player was replaced with dog meshes that were available in the game. The game features a total of 8 canine breeds, each varying in shape, size and texture (Fig. 1a,b). Some of these breeds are adorned with unique collars, adding an extra layer of distinction to their appearance. These breeds include Labrador, Rottweiler, Shepherd, Wolf, Coyote, Poodle, Terrier, and Pug. The dogs contain an extensive library of animations such as walking, cantering, running, sitting and barking, among others. Our dataset comprises 118 videos, totaling 27900 frames, capturing a diverse range of scenarios. We aimed to maintain an equal distribution of frames for each dog type (Fig. 1c). For the training and testing phases, we specifically selected 6 dogs and 2 dogs, respectively, maintaining approximately an 80:20 ratio. In each video, the dogs performed either a single animation, like running, or a sequence of animations, such as walking followed by sitting. This explains the elevated value of the walking bar in the bar chart (Fig. 1d). Alongside the RGB images, we generated depth maps, kinematic skeletal motion sequences, 2D/3D keypoint coordinates, segmentation maps and camera intrinsics and extrinsics. The 3D keypoints were extracted from the digital skeleton also referred to as a rig. We extracted the skeletal information from the following 26 joints: head, neck, throat, mid-spine, pelvis, start-tail, mid-tail, end-tail, five joints in each front leg, and four joints in each back leg. The facial keypoints were omitted because the facial joints in the rig did not align accurately with the joints in image space. Accurate detection of body parts using the RGB image is important for the network; misalignment may result in erroneous joint detection, leading to incorrect predictions of 3D joints.

	Outdoor	# Dogs	# Keypoints	# Images
DigiDogs (ours)	✔	8	33	27900
RGBD-Dog [14]	✗	7	41	136,000
SVM	✗	51	29	1880

Table 1. Comparison of DigiDogs (Ours) with real-world indoor 3D dog pose datasets.

4. D-Pose architecture and losses

To achieve 3D dog pose from a single image, we extend the DINOv2 [22] model by adding 3 pose branches simultaneously predicting the joint heatmaps for each X, Y, Z axis in the XY, XZ, ZY planes, respectively. The heatmaps are marginal heatmaps representing the likelihood of the joint positions in an image [21]. To generate the ground truth heatmaps, the joint positions were first normalised within the Normalized Device Coordinate space, ensuring they fell within the range of [-1, 1]. Subsequently, these normalised positions were then represented as a 2D heatmap using a Gaussian distribution. The DINOv2 model

was used rather than a conventional pose estimation model such as the stacked hourglass network due to the latter's superior domain adaptation performance [22]. We refer to our customised model as D-Pose for ease of reference and distinction (Fig. 1e). D-Pose's input is an RGB image of size $448\times448\times3$ and outputs 3 heatmaps of size $32\times32\times K$ where K for the number of keypoints. H_{xy}, H_{zy}, and H_{xz} denote the predicted heatmaps for the heatmaps on the xy, zy, and xz planes, respectively. We denote the ground truth heatmaps with a hat symbol: $\hat{\cdot}$. Within each branch, a linear layer is introduced to reduce the feature dimension from 768 to 512, followed by a convolutional layer with a kernel size set to 1. To predict the H_{zy} and H_{xz} heatmaps, an axis permutation technique [21] was applied following the dimension reduction operation. This axis permutation entails transposing the intermediary activations, effecting a seamless transition from the xy space to the zy and xz spaces. To refine the network's performance, we fine-tune it by unfreezing the last three layers and introduce a dropout layer just before the final layer, with a probability of 0.3 for zeroing out elements.

We have adapted the loss in [21] by masking elements of the loss function based on the availability/visibility of the joints. The Jensen-Shannon divergence (JSD) [17] is computed between the predicted and ground truth heatmaps to encourage them to mimic the shape of a specific probability distribution. Additionally, the mean squared error between predicted and actual joint locations is calculated. The following equation Eq. (1) shows the loss used for training:

$$\begin{aligned}\mathcal{L}_1 = v \cdot \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|_2 + \\ \text{JSD}(H_{xy} \| \hat{H}_{xy}) + \\ \text{JSD}(H_{zy} \| \hat{H}_{zy}) + \\ \text{JSD}(H_{xz} \| \hat{H}_{xz})\end{aligned} \tag{1}$$

Where v is a binary representation of joint availability, with a length of K corresponding to the number of keypoints. $\boldsymbol{\mu}$ represents the joint coordinates and is computed the following way:

$$\begin{aligned}x_{xy}, y_{xy} &= \mathbb{E}[H_{xy}] \\ y_{zy}, z_{zy} &= \mathbb{E}[H_{zy}] \\ x_{xz}, x_{xz} &= \mathbb{E}[H_{xz}] \\ \boldsymbol{\mu} &= (x_{xy}, y_{xy}, \frac{z_{zy} + z_{xz}}{2})\end{aligned} \tag{2}$$

We set the predictions from the $\hat{H}_{xy}$ for the xy-coordinates. However, we could have derived the xy-coordinates from the zy- and xz-heatmaps. To obtain the z-coordinate, we averaged the extracted z-coordinate from H_{zy} and H_{xz}. The JSD function's components were selectively masked according to Eq. (3), with the binary mask $M(x, y)$ determining inclusion ($M(x, y) = 1$) or exclusion ($M(x, y) = 0$) from the calculation.

$$\text{JSD}(H \| \hat{H}) = \frac{1}{2}[D_{KL}(H \| M \cdot \hat{H}) + D_{KL}(\hat{H} \| M \cdot H)] \tag{3}$$

5. Experiments

We present the implementation details (Sec. 5.1), the test datasets employed for thorough evaluation (Sec. 5.2), and the evaluation protocols (Subsection Sec. 5.3) adopted in our research. We showcase both qualitative and quantitative results (Sec. 5.4) obtained from both synthetic and real images, providing a comprehensive analysis of our findings.

5.1. Implementation details

D-Pose's architecture was built upon the DINOv2 backbone, employing a ViT-B [8] with a patch size of 14. The network is trained using PyTorch Lightning [10], with a batch size set to 16 and a maximum number of epochs set to 500. We selected the Adam optimizer with an initial learning rate of 1e-5 for optimization. The learning rate was adjusted according to the multi-step learning rate scheduler, by decreasing the learning rate every 5 epochs with γ set to 0.001. Additionally, to prevent overfitting and to ensure efficient training, we employed early stopping after 5 epochs. Data augmentation techniques were applied to augment the dataset, including random color jitter (p=0.45), Gaussian blur (p=0.35), and random grayscale conversion (p=0.5). p is the probability for randomly applying the augmentations. Furthermore we randomly cropped the image based on the bounding box, note that by doing this the center point (c_x, c_y) of the camera changes.

5.2. Real-world dog datasets for evaluation

We evaluate our model on the DigiDogs test set, which are dogs/scenes that the model has not seen during training. Furthermore, our model is evaluated on real world datasets which we will further discuss in this section.

StanfordExtra (2D) [4] Contains 12k in-the-wild images of dogs including 2D ground truth such as keypoint coordinates and segmentation maps. We exclude images labeled with multiple dogs and those lacking segmentation maps.

RGBD-Dog (3D) [14] Contains multi-view videos of dogs in mocap suits recorded in a controlled environment. This dataset is used to evaluate our 3D pose estimation and the generalisation capability of our network. Furthermore, this dataset was also used to train a network using only real-world data and subsequently compare its performance with the network trained solely on synthetic data (DigiDogs). We refer the reader to **Protocol 2** in Sec. 5.3.

SVM is an internal dataset collected specifically for detecting different kinds of lameness using diverse sensor and

capture systems, including pressure mats, optical markers, IMUs, and video/depth cameras. The capture included 64 subjects, each performing three walking and three trotting trials. Eight synchronised RGBD cameras were used to record the data alongside the standard systems in a laboratory. Due to the nature of the data and ongoing research, we have decided to use it for the current study without formally contributing it as a standalone dataset. Due to the capture setup, we plan to further expand its use and aim to release it for public access in future work. We employ a subset of the SVM dataset to assess the performance of our network. However, it is essential to acknowledge that the depth information is available solely from one side of the dog.

5.3. Evaluation protocols

To assess our network's performance, the percentage of correct keypoints (PCK) and the mean per joint position error (MPJPE) were used. PCK measures the percentage of predicted joint locations that fall within a specific threshold α of the ground truth keypoint locations. On the other hand, MPJPE quantifies the distance between the predicted and ground truth locations. We report these metrics for both 2D and 3D scenarios. When computing the metrics, we only consider the joints that are visible and have ground truth labels available. The PCK and MPJPE in 3D are measured after Procrustes alignment of the predicted and ground truth skeletons. Due to the lack of facial keypoints in DigiDogs, the metric calculations apply exclusively to the remaining anatomical components, excluding the facial region.

Protocol 1 (2D): Demonstrating the 2D re-projection errors. Following previous works [4,24], we evaluate on StanfordExtra test set in terms of 2D reprojection error. For PCK evaluation, we set the PCK threshold to $\alpha = 0.15$. The distances between predicted and ground truth joint locations are normalised relative to the area of the 2D segmentation map. Our method is compared to the results against recent state-of-the-art methods such as WLDO and BARC [4, 24], both of which leverage the animal parametric model for 3D pose estimation.

Protocol 2 (3D): Comparing real to synthetic data. In this protocol, our approach is trained with real-world indoor 3D pose datasets. This protocol presents the 3D pose estimation results. We compare D-Pose trained on the DigiDogs dataset with D-Pose trained on real-world datasets (RGBD-Dog [14] and SVM). We evaluate our results on different datasets such as the StanfordExtra (2D), the RGBD-Dog (3D) and the SVM dataset (3D). We set the threshold to $\alpha = 0.15$ and normalise the skeleton with respect to the length between the neck and pelvis for the 3D metrics. For future reference, we identify D-Pose, the pose estimation model, based on the specific dataset it is trained on (Tab. 2).

Symbol	Training dataset
D-Pose$_{digi}$	DigiDogs (synthetic)
D-Pose$_{rgbd}$	RGBD-Dog (real & indoor)
D-Pose$_{svm}$	SVM dataset (real & indoor)

Table 2. The nomenclature explains the symbols and specifies the dataset used for training the network.

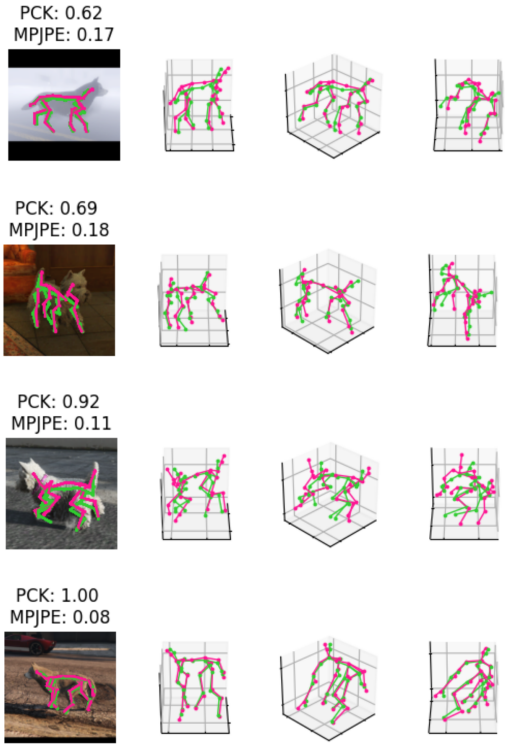

Figure 2. 3D results (PCK and MPJPE) from D-Pose$_{digi}$ on the DigiDogs test set. The ground truth and predicted skeleton are coloured green and pink, respectively.

5.4. Results

We begin by presenting both quantitative and qualitative results on the DigiDogs test set. D-Pose$_{Digi}$ achieves a PCK$_{2D}$ of 95.28 and a MPJP$_{2D}$ of 0.06, along with PCK$_{3D}$ of 78.27 and MPJPE$_{3D}$ of 0.14. Fig. 2 shows the 3D metrics including D-Pose$_{Digi}$'s ability to generate high quality 3D poses on the in-domain test set. The network was not exposed to these specific dog types or background environments during training.

Protocol 1 (2D): We compare our method to WLDO [4] and BARC [24], which rely on parametric models and 2D supervision. Tab. 3 presents the 2D results, while Fig. 3 provides qualitative results of our method on the StanExt test set. Our results using the 2D metrics are better than the current state-of-the-art methods, without being trained on the StanfordExtra training set or using real-world data and/or a

Figure 3. 2D results (PCK, MPJPE) on the StanfordExtra test set from D-Pose$_{digi}$ (left) and from the D-Pose$_{rgbd}$ (right). The 2D projections including the predicted 3D pose from different angles are shown. The ground truth and predicted skeleton are coloured green and pink, respectively. The ground truth skeleton is only shown in 2D as the StanfordExtra is a 2D dataset. Furthermore, only the visible joints are shown from the ground truth skeleton.

geometric prior. These results is most likely due to the use of 3D ground truth, in contrast to exclusively relying on 2D ground truth. We believe that using our approach in conjunction with a geometric prior has the potential to enhance the performance of existing 3D dog pose estimation methods. As illustrated in Fig. 3, our D-Pose$_{Digi}$ model demonstrates the ability to reconstruct both 2D and 3D dog poses from real-world images, despite being trained on synthetic 3D data. It is important to note that in Fig. 3, we present the ground truth skeleton (coloured in green) in 2D, as the StanfordExtra dataset only provides 2D ground truth annotations.

Protocol 2 (3D) In Tab. 3, we present the 2D reprojection results obtained from D-Pose$_{rgbd}$, a model trained exclusively on real-world indoor data (RGBD-Dog) and tested on the StanfordExtra dataset. Despite the precise 3D pose labels in the RGBD-Dog dataset, its lack of diversity becomes evident. This limitation is why D-Pose$_{digi}$ outperforms D-Pose$_{rgbd}$ by 12% in the 2D pose estimation task. The table clearly illustrates the substantial advantage of leveraging DINOv2 in bridging the domain gap. Notably, D-Pose$_{rgbd}$, despite not being trained on the StanfordExtra dataset, demonstrates performance levels nearing the state

Method	PCK@0.15		
	Avg	Legs	Tail
WLDO [4]	78.8	76.4	63.9
BARC [25]	82.8	82.3	63.3
Ours (D-Pose$_{rgbd}$)	71.8	72.0	**86.6**
Ours (D-Pose$_{digi}$)	**83.8**	**86.1**	80.0

Table 3. 2D results on the StanfordExtra test set [4]. Comparison to *SOTA*. Results from WLDO and BARC are reproduced from [26]. *Our(s)* networks are trained either on DigiDogs or RGBD-Dog datasets.

of the art. In summary, it remains crucial to diversify the training dataset for more robust pose reconstruction, especially when dealing with images captured in natural environments (in-the-wild). This also becomes evident when qualitatively comparing the 3D skeleton predictions in row 2 of Fig. 3. The predictions generated by D-Pose$_{digi}$ appear more plausible than those produced by D-Pose$_{rgbd}$. This is most likely attributed to the broader range of poses, including sitting and lying down, and the diverse background scenes featured in the DigiDogs dataset. In contrast, although the RGBD-Dog dataset is larger Tab. 1, its limited background variety adversely impacts the overall quality of pose predictions. This shows that using synthetic training data offers more advantages than capturing 3D data in a controlled laboratory environment. Synthetic data is not only easier to generate but also provides a greater diversity of scenarios and poses for training purposes. As there are no available in-the-wild 3D pose datasets, our method, using training synthetic data is a viable alternative to estimate the 3D pose of dogs from images.

We continue to present the results obtained from the D-Pose$_{rgbd}$ model, specifically evaluating its 3D performance on the RGBD-Dog test sets. There are two RGBD-Dog test sets: one assessing the model's adaptability to missing data, while the other evaluates its ability to generalise to previously unseen dog breeds. Furthermore, we perform a comparative analysis by contrasting it with the D-Pose$_{digi}$ model. Tab. 4 demonstrates that D-Pose$_{rgbd}$ is able to predict the 3D poses for test sets encompassing dogs 1 to 5, which it has encountered during the training process with an acceptable level of accuracy. Tab. 4 also demonstrates D-Pose$_{rgbd}$'s generalisation capability by successfully predicting poses for previously unseen dogs, specifically dog 6 and 7. When compared to D-Pose$_{rgbd}$, D-Pose$_{digi}$ appears to struggle in accurately predicting poses across the test sets. We believe this is due to the domain gap in the z-coordinate space, due to the different z-coordinate distributions across different datasets, which can be caused by different camera setups. However, upon qualitative evaluation of the results as depicted in Fig. 4, D-Pose$_{digi}$ pose

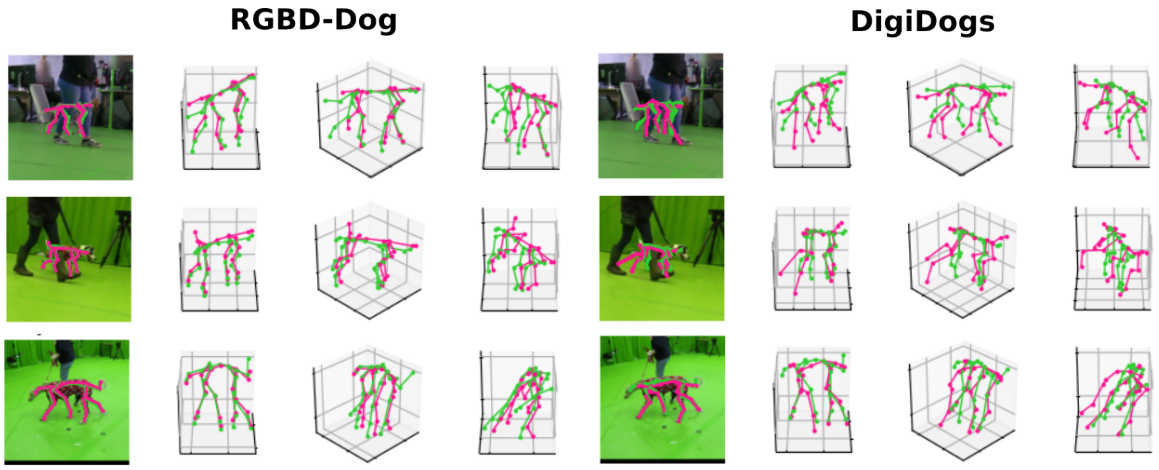

Figure 4. Qualitative results from D-Pose$_{rgbd}$ (left) vs. D-Pose$_{digi}$ (right). The ground truth and the predicted skeletons are coloured in green and pink, respectively.

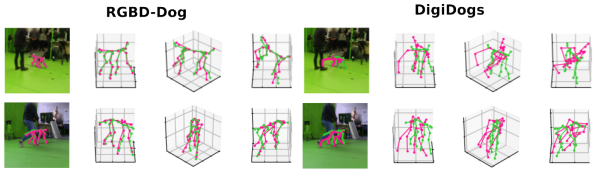

Figure 5. Failure cases from D-Pose$_{digi}$ compared to successful cases from D-Pose$_{rgbd}$ (from left to right). The ground truth and predicted skeletons are coloured in green and pink, respectively.

predictions appear plausible. D-Pose$_{digi}$ faces challenges in accurately predicting 3D poses, especially in cases of occlusion or when a human is present in the frame. It often misidentifies human limbs as dog limbs, leading to confusion in pose predictions (Fig. 5). Overall, the qualitative results show potential in our method, as it is important to reiterate that the model has not been exposed to the RGBD-Dog data during its training process.

Finally, we present the 3D results on the SVM test set from the D-Pose$_{svm}$, D-Pose$_{rgbd}$ and D-Pose$_{digi}$ models. D-Pose$_{svm}$ achieves high accuracy on the SVM dataset, which belongs to the same domain. While D-Pose$_{svm}$, performs well on its in domain dataset, D-Pose$_{rgbd}$ and D-Pose$_{digi}$ do not perform well on the out-of-domain test set quantitatively (Tab. 5). Similarly to the previous analysis of D-Pose$_{rgbd}$, we believe this is due to the domain gap in the z-coordinate space. Nonetheless, upon closer examination, we find that the predicted 3D poses exhibit a reasonable proximity to the ground truth, as illustrated in Fig. 6. Despite narrowing the domain gap in 2D image space, differences in Z-coordinate distributions persist across the 3D datasets due to factors like varying camera setups. A domain gap in the z-axis still remains. We believe that this could be addressed by refining the 3D pose using label refinement networks [31]. Furthermore, while the SVM dataset only includes the 3D ground truth from the lateral side, D-Pose$_{rgbd}$ and D-Pose$_{digi}$ were trained with datasets including complete skeleton 3D ground truth data. Because of this, both models are able to predict believable 3D skeletons (Fig. 7). It should be noted that D-Pose$_{digi}$ consistently exhibits a more reliable 3D skeletal pose structure when observed from different angles, compared to the predictions made by D-Pose$_{rgbd}$

6. Conclusion

We present an approach for automatically extracting the 3D pose of dogs from single-view RGB images, using synthetic training data. Past methodologies relied on 2D weakly supervised techniques due to the absence of suitable 3D datasets, yielding promising results but leaving persisting challenges in depth perception. To overcome these depth ambiguities, we created a synthetic 3D pose dataset, DigiDogs, through modifications to the video game Grand Theft Auto. Moreover, to help bridge the gap between synthetic and real-world data, we harnessed the generalisation capabilities of the DINOv2 foundation model and fine-tuned it for the 3D pose estimation task. While previous research has primarily assessed their approach using 2D metrics, we extend our analysis by providing results in 3D as well. Firstly, we demonstrate that training our architecture with our synthetic dataset, DigiDogs, yields significantly more realistic and comprehensive 3D poses compared to training it on real-world indoor datasets. This is due to the dataset's diverse range in of dog appearances, poses and contextual scenes. Secondly, we outperform the current state-of-the-art in 2D. Finally, through a comprehensive blend of qualitative and quantitative analyses, we have established the practicality of estimating realistic 3D

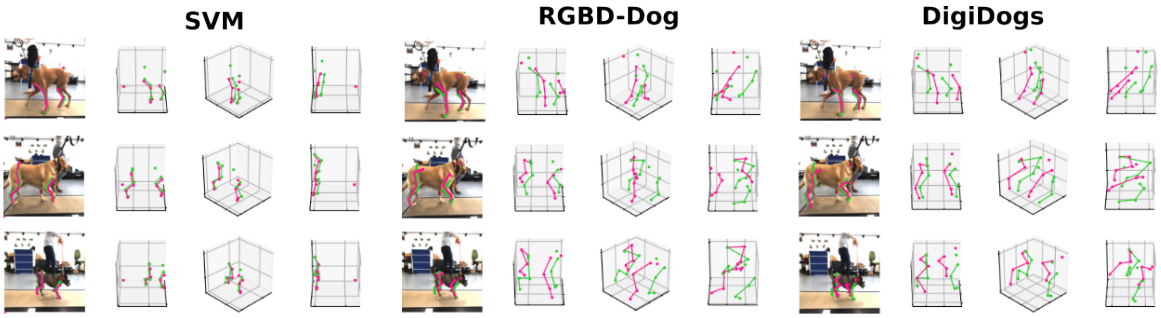

Figure 6. Qualitative results on the SVM test set from D-Pose$_{svm,rgbd,digi}$ (from left to right). Ground truth and predicted skeletons are coloured in green and pink, respectively.

Dog	Training set	Metric	Avg
Dog1	RGBD-Dog	PCK	75.23
		MPJPE	0.16
	DigiDogs	PCK	31.11
		MPJPE	0.44
Dog2	RGBD-Dog	PCK	78.33
		MPJPE	0.15
	DigiDogs	PCK	29.63
		MPJPE	0.39
Dog3	RGBD-Dog	PCK	73.12
		MPJPE	0.19
	DigiDogs	PCK	29.01
		MPJPE	0.48
Dog4	RGBD-Dog	PCK	73.54
		MPJPE	0.16
	DigiDogs	PCK	30.25
		MPJPE	0.38
Dog5	RGBD-Dog	PCK	67.01
		MPJPE	0.29
	DigiDogs	PCK	30.35
		MPJPE	0.44
Dog6	RGBD-Dog	PCK	76.01
		MPJPE	0.16
	DigiDogs	PCK	29.87
		MPJPE	0.39
Dog7	RGBD-Dog	PCK	75.13
		MPJPE	0.18
	DigiDogs	PCK	30.83
		MPJPE	0.47

Table 4. 3D results (PCK and MPJPE) on the RGBD-Dog dataset [14] from D-Pose$_{rgbd}$ and D-Pose$_{digi}$. Dog 6 and Dog 7 were not seen by the network.

poses of dogs from real-world images, using a combination of the DINOv2 model and synthetic training data.

Limitations and Future work Our method reconstructs plausible 3D skeleton structures of dogs from real-world

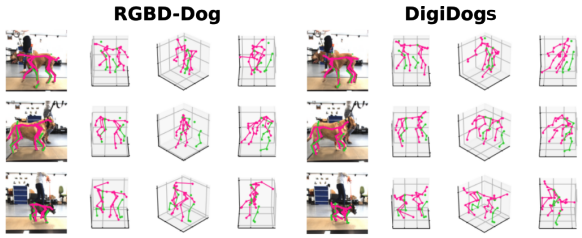

Figure 7. Qualitative results on the SVM test from D-Pose$_{rgbd}$ and D-Pose$_{digi}$. Demonstrating complete results as the SVM dataset only consists of joints from the lateral side. Ground truth and predicted skeletons are coloured in green and pink, respectively.

Method	PCK	MPJPE
D-Pose$_{svm}$	86.15	0.31
D-Pose$_{rgbd}$	20.34	0.40
D-Pose$_{digi}$	21.35	0.44

Table 5. 3D results (PCK and MPJPE) on the SVM test set from networks trained on the following datasets: SVM, RGBD-Dog and DigiDogs.

images qualitatively. However, quantitative results reveal a domain gap in the z-coordinate space, indicating room for improvement. We plan to refine the joint predictions [32] to match specific 3D pose datasets. After refining the joints, we are optimistic that by integrating our methodology with a geometric prior could enhance existing 3D dog pose estimation techniques which relied solely on 2D ground truth. In addition, we plan to evaluate the significance of diversity, including various dog breeds and textures, and determine the optimal size of the training synthetic data to enhance overall performance.

Acknowledgments This research was supported by EPSRC Audio-Visual Media Platform Grant EP/P022529/1 and by the Leverhulme Trust Early Career Fellowship scheme.

References

[1] Grand theft auto v, [2014]. New York, NY :Rockstar Games. 1

[2] M. Badger, Yufu Wang, Adarsh Modh, Ammon Perkes, Nikos Kolotouros, Bernd Pfrommer, Marc F Schmidt, and Kostas Daniilidis. 3d bird reconstruction: a dataset, model, and shape recovery from a single view. *Computer vision - ECCV ... : ... European Conference on Computer Vision : proceedings. European Conference on Computer Vision*, 12363:1–17, 2020. 2

[3] Praneet C. Bala, Benjamin R. Eisenreich, Seng Bum Michael Yoo, Benjamin Y. Hayden, Hyun Soo Park, and Jan Zimmermann. Openmonkeystudio: Automated markerless pose estimation in freely moving macaques. *bioRxiv*, 2020. 2

[4] Benjamin Biggs, Ollie 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. 1, 2, 4, 5, 6

[5] Benjamin Biggs, Thomas Roddick, Andrew Fitzgibbon, and Roberto Cipolla. Creatures great and SMAL: Recovering the shape and motion of animals from video. In *ACCV*, 2018. 3

[6] 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. 3

[7] Jake Deane, Sinead Kearney, Kwang In Kim, and Darren Cosker. Dynadog+t: A parametric animal model for synthetic canine image generation. *CoRR*, abs/2107.07330, 2021. 2

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 4

[9] Matteo Fabbri, Guillem Brasó, Gianluca Maugeri, Aljoša Ošep, Riccardo Gasparini, Orcun Cetintas, Simone Calderara, Laura Leal-Taixé, and Rita Cucchiara. Motsynth: How can synthetic data help pedestrian detection and tracking? In *International Conference on Computer Vision (ICCV)*, 2021. 3

[10] William Falcon and The PyTorch Lightning team. PyTorch Lightning, Mar. 2019. 4

[11] Le Jiang, Shuangjun Liu, Xiangyu Bai, and Sarah Ostadabbas. Prior-aware synthetic data to the rescue: Animal pose estimation with very limited real data. In *The British Machine Vision Conference (BMVC)*, 11 2022. 1

[12] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 1

[13] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher. Dreampose: Fashion image-to-video synthesis via stable diffusion, 04 2023. 3

[14] Sinead Kearney, Wenbin Li, Martin Parsons, Kwang In Kim, and Darren Cosker. Rgbd-dog: Predicting canine pose from rgbd sensors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 3, 4, 5, 8

[15] S. Lee, H. Park, D. Kim, J. Kim, M. Boboev, and S. Baek. Image-free domain generalization via clip for 3d hand pose estimation. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2933–2943, Los Alamitos, CA, USA, jan 2023. IEEE Computer Society. 3

[16] Ci Li, Nima Ghorbani, Sofia Broomé, Maheen Rashid, Michael J. Black, Elin Hernlund, Hedvig Kjellström, and Silvia Zuffi. hsmal: Detailed horse shape and pose reconstruction for motion pattern recognition. *CoRR*, abs/2106.10102, 2021. 2

[17] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. 4

[18] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 2

[19] Jesse D. Marshall, Ugne Klibaite, Amanda Gellis, Diego E. Aldarondo, Bence P. Ölveczky, and Timothy W. Dunn. The pair-r24m dataset for multi-animal 3d pose estimation. *bioRxiv*, 2021. 2

[20] Jiteng Mu, Weichao Qiu, Gregory D. Hager, and Alan L. Yuille. Learning from synthetic animals. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 3

[21] Aiden Nibali, Zhen He, Stuart Morgan, and Luke A. Prendergast. 3d human pose estimation with 2d marginal heatmaps. *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1477–1485, 2018. 3, 4

[22] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 2, 3, 4

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. 3

[24] Nadine Rueegg, Silvia Zuffi, Konrad Schindler, and Michael J. Black. BARC: Learning to regress 3D dog shape from images by exploiting breed information. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3876–3884, 2022. 1, 2, 5

[25] Nadine Rueegg, Silvia Zuffi, Konrad Schindler, and Michael J. Black. Barc: Breed-augmented regression using classification for 3d dog reconstruction from images. *International Journal of Computer Vision*, 131(8):1964–1979, Aug 2023. 6

[26] Nadine Rüegg, Shashank Tripathi, Konrad Schindler, Michael J. Black, and Silvia Zuffi. BITE: Beyond priors

for improved three-D dog pose estimation. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8867–8876, June 2023. 2, 6

[27] Moira Shooter, Charles Malleson, and Adrian Hilton. Sydog: A synthetic dog dataset for improved 2d pose estimation, 2021. 3

[28] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 358–374. Springer, 2022. 3

[29] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 3

[30] Yufu Wang, Nikos Kolotouros, Kostas Daniilidis, and M. Badger. Birds of a feather: Capturing avian shape models from images. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14734–14744, 2021. 2

[31] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J. Cashman, and Jamie Shotton. Fake it till you make it: Face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3681–3691, October 2021. 7

[32] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Matthew Johnson, Virginia Estellers, Thomas J. Cashman, and Jamie Shotton. Fake it till you make it: Face analysis in the wild using synthetic data alone, 2021. 3, 8

[33] 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 *International Conference on Computer Vision*, Oct. 2019. 3

[34] Silvia Zuffi, Angjoo Kanazawa, David W. Jacobs, and Michael J. Black. 3d menagerie: Modeling the 3d shape and pose of animals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2