ELSEVIER

Computer Vision and Image Understanding
journal homepage: www.elsevier.com

Animal Pose Estimation: A Closer Look at the State-of-the-Art, Existing Gaps and Opportunities

Le Jiang[a], Caleb Lee[a], Divyang Teotia[a], Sarah Ostadabbas[a,**]

[a]*Augmented Cognition Lab, Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA.*

ABSTRACT

Over the past few years, research on animal pose estimation in computer vision field has grown in many aspects such as 2D and 3D pose estimation, 3D mesh reconstruction, and behavior prediction. Promoted by deep learning, more and more animal pose estimation tools and animal pose datasets have also been made publicly available. However, compared to human pose estimation, which already has high accuracy and high applicability for complex scenes, animal pose estimation is still at a preliminary stage. The huge domain shift between each species, the scarce datasets, and uncooperative research subjects all pose intractable challenges to the development of robust and accurate animal pose estimation algorithms. In this review paper, we summarize the recent (from 2013 to 2021) work in animal pose estimation from computer vision perspective in order to present the state-of-the-art approaches and highlight the challenges they face in this field. We first categorize the various methods of animal pose estimation and present them according to several keywords. Also, we sort and introduce the released annotated image, video, and 3D models of animal poses as well as a promising substitute for real dataset. We also report the performances of the existing algorithms and visualize their results. Finally, we provide an in-depth analysis of the persisting obstacles in this field based on existing work, and offer potential solutions.

© 2024 Elsevier Ltd. All rights reserved.

1. Introduction

Pose estimation in the computer vision field is the specific task of localizing the joint positions or predefined keypoints of an object (e.g. a car, a human body, or a dog) in an image. It goes back to the early works of the 1990s, which were aimed at human detection, motion tracking, and facial pattern estimation (Yang and Huang, 1994; Sung and Poggio, 1998). Over the last decade, the cost and technical difficulties for pose estimation have become considerably lower. As such, its applications have gradually been integrated into all aspects of our lives, including (1) gesture-based human-computer interaction (Nguyen et al., 2020), (2) assessment and correction of human movement and posture in healthcare and sport applications (Chen and Yang, 2020; Chen et al., 2018a), (3) social security – detection of adversary actions (Tsiktsiris et al., 2020), and (4) gaming and 3D avatar generation in virtual/augmented reality environments (Obdrzalek et al., 2012), among others. This remarkable progress is attributed to the introduction of deep learning, which has led to an explosive increase in pose estimation works, opening the door to other branches of this topic including animal pose estimation.

Animal pose estimation plays an essential role in learning and understanding animal behavior (Anderson and Donath, 1990; Butail et al., 2015; Del Pero et al., 2015a; Joska et al., 2021), preserving animal's appearance information (Duncan et al., 2017), protecting endangered species (Zuffi et al., 2019), understanding the migration of wild animals (Li et al., 2014; Bauer and Klaassen, 2013), and even taking care of our pets (Biggs et al., 2020). However, animal pose estimation has many challenges not present in the human pose estimation problem. In terms of public attention, it has little influence on human-specific healthcare (Huang et al., 2021; Liu et al., 2022) or military tasks, making it more difficult to secure funding. Until now, animal keypoint detection tasks are still rarely seen in the image processing challenge held at international computer vision

[**] Corresponding author: Tel.: +1-617-373-4992;
e-mail: ostadabbas@ece.neu.edu (Sarah Ostadabbas)

conferences. In terms of data acquisition, most live animals are more uncooperative than humans, so collecting rich sets of pose data from them is burdensome. Researchers can conduct 3D scanning (Duncan et al., 2017) and take photos (Russello et al., 2021; Mathis et al., 2021) of domesticated animals, such as cow and horses on farms and in zoos. People can even put dogs in the motion capture suit (Ricardo, 2022) to collect the precise 3D pose. However, for lions and tigers with high aggression, or for elephants and giraffes with large bodies, it is unlikely to make them stay in the monitoring spot obediently. In works such as Joska et al. (2021), high-speed cameras are set up to collect the data of several raised cheetahs in the wild; however, these cases are difficult to replicate, especially for the endangered species (Zuffi et al., 2019) that should be interfered with. In addition, obtaining data by tracking animals in the wild requires tons of manpower, materials, and funds, which at the end could be still very limited due to the adverse vision conditions such as occlusions and huge illumination changes. While there are many publicly available large-scale human pose datasets, such as MPII human pose (Andriluka et al., 2014a), COCO (Lin et al., 2014b), and SMPL (Loper et al., 2015), the scarcity of the annotated animal pose datasets hinders the use of well-developed human pose/shape estimation model structures in supervised fashion. In terms of model, human and animal pose estimation can share a similar structure and backbone. However, the design for animal pose estimation models must take into account data scarcity and the diversity of animal species. Under the same network, the human body can achieve high-precision pose estimation since there are many large-scale human pose datasets available. In contrast, animal pose estimation models are suffering from both data scarcity and lack of model variety. Furthermore, the huge differences in physical characteristics among animals species cause a large domain discrepancy, as a network trained with images of cats will not be generalized to images of giraffes due to differences in their bone structure, shapes, and textures. Although some works (Mathis et al., 2021; Yu et al., 2021a; Cao et al., 2019; Li and Lee, 2021) have confirmed that it is possible to train models with generalization ability on the unseen animal, the accuracy obtained on the out-domain data is obviously lower than that obtained on in-domains data.

Despite the difficulties throughout the development of animal pose estimation algorithms, many gratifying results have been achieved in last few years. As listed in Table 1, these methods can be categorized into 2D pose estimation, 3D pose estimation, model reconstruction and behavior prediction, based on the objectives of the algorithms. In the 2D and 3D pose estimation, Cao et al. (2019) analyzed and mitigated the domain shift between human and animals and among different species, and fine-tuned a human pose estimator, AlphaPose (Hao-Shu Fang and Cewu Lu, 2017), for animal pose estimation. Meanwhile, Li and Lee (2021) and Mu et al. (2020) trained their model with the combination of synthetic data and real data and designed a domain adaptation module to reduce the synthetic vs. real domain gaps. There are also a series of end-to-end pose estimation toolboxes (Graving et al., 2019a; Mathis et al., 2018), which are based on deep neural networks and their derivative works (Nath et al., 2019; Lauer et al., 2021). The introduction of Skinned Multi-Animal Linear (SMAL) model in (Zuffi et al., 2017) has opened up the field of reconstructing animal models from a single image (Zuffi et al., 2018, 2019; Biggs et al., 2018b, 2020; Kanazawa et al., 2018a). SMAL greatly reduces the difficulty of modeling endangered species in the wild. Also, it provides an effective way to lifting 2D to 3D poses.

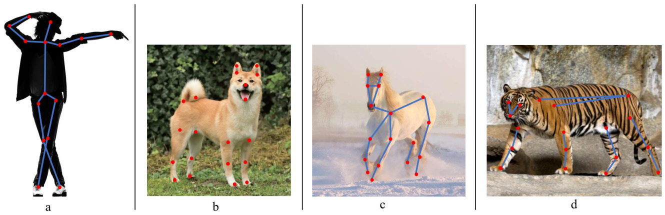

Fig. 1: Different keypoints and skeleton annotations on human and animals, where (a) shows a standard for human pose annotation used in (Andriluka et al., 2014b), while three types of annotation for animal pose are from (b) (Biggs et al., 2020), (c) (Cao et al., 2019), and (d) (Pero et al., 2016). The red points mark the keypoint on the image and the blue lines which connect the keypoints denote the skeleton.

In this review paper, in order to provide an informative survey in the topic of animal pose estimation, we first need to have a comprehensive understanding of the capabilities and drawbacks of the current state-of-the-art (SOTA). This will allow us to obtain inspiration, identify existing challenges, and plan for future directions in this field. This review paper presents an overview of the SOTA animal pose estimation work and makes the following contributions:

- Collecting and categorizing animal pose estimation work from 2013 to 2021 into 2D and 3D pose estimation, model reconstruction, and behavior prediction, focusing mainly on the larger quadruped mammals, as listed in Table 1.
- Sorting and reporting publicly available animal pose datasets based on their animal classes, data types, annotation types, and data sizes as listed in Table 2.
- Summarizing the quantitative evaluation of the SOTA animal pose estimation works.
- Discussing the challenges and remaining gaps in the field, and exploring the existing opportunities as well as offering potential solutions/next steps.

The rest of the paper is organized as follows. Section 2 provides a comparison between the pose-based keypoint definition of human vs. animal. In Section 3, we describe the SOTA animal pose estimation methods categorized based on their objectives. We report the publicly available animal pose datasets in Section 4. In Section 5, we compare the results of the SOTA animal pose estimation models. Finally in Section 6 and Section 7, we conclude this survey by discussing the existing gaps and opportunities.

2. Animal Pose Definition

Animal or human pose estimation is about localizing their body joints. In this section, we describe the differences between the definition of joints in human vs. animals. Some examples of these differences can be seen in Fig. 1. In Fig. 1(a), the keypoints and skeleton of a human subject in the MPII dataset (Andriluka et al., 2014b) is shown. Compared to the MPII, which is mainly used for human action detection, COCO dataset (Lin et al., 2014b), another human pose dataset, pays more attention to human facial features. Thus, it has more keypoints on the eyes, nose and ears than those in the MPII. In Fig. 1(b), Biggs et al. (2020) chose to set 2 keypoints on each ear and tail and no keypoints on the eyes. In Fig. 1(d), the TigDog dataset (Pero et al., 2016) focuses more on the body posture of the animal, so there is no landmark on the ear and tail. By observing the distribution of animal keypoints among the three works in Fig. 1(b)-(d) (Andriluka et al., 2014b; Biggs et al., 2020; Cao et al., 2019), it is apparent that while the objectives of these studies could be different, the definition and distribution of the keypoints greatly overlap among quadrupeds. Around 21 keypoints are annotated for these larger mammals including two eyes, one nose, one neck, two ears, four ankles, four knees, four elbows, and one tail-base.

In many cases, the keypoint distribution in the large-scale human pose datasets is similar to those of the larger-in-size quadruped animals, which is due to the morphological similarity between humans and quadruped mammals. Recent work attempted and succeeded in proving that the existing dense pose recognition (Güler et al., 2018), detection and segmentation can be transferred between human and animal poses which is possible for some animals that are physically similar to human, such as chimpanzees (Sanakoyeu et al., 2020). For example, Neverova et al. (2020) transferred information between humans and quadrupeds such as cats, dogs, elephants, giraffes, horses, bears, etc., by using functional maps (Ovsjanikov et al., 2012) to relate different 3D shape. This extended the field of cross-domain adaptation to more species of animals. However, there are many animals whose pose estimation face serious domain shifts when relying on human-specific pose models. Cao et al. (2019) pointed out that the differences in defined "bones" between each keypoint would also cause a great domain shift even among similar species, let alone between human and animals. The relative length of the defined "bones" of the human is shorter in upper body, while the relative length is longer in lower body than that of the quadruped. Animals with special physical characteristics, such as giraffes, will have more serious domain shifts. Therefore, solving this domain shift will greatly alleviate the data scarcity problem faced by animal pose estimation.

3. Animal Pose Estimation Models

The human pose estimation has been developed for decades, and the usage of deep learning and convolutional neural networks since 2014 has made many major breakthroughs in the computer vision field. Meanwhile, the animal pose estimation topic that has been gradually developed in the past ten years, naturally refers to many algorithms developed for human pose estimation. Taking advantage of this history, we classify animal pose estimation methods into three main categories of: (1) 2D animal pose estimation, (2) 3D animal pose estimation, and (3) 3D animal mesh recovery. The categorization of animal pose estimation works from 2013 to 2021 is summarized in Table 1, while a generic pipeline for such works in demonstrated Fig. 2.

3.1. 2D Animal Pose Estimation

2D pose estimation aims to detect the 2D coordinates of the keypoints (joints) of an animal in a single image or a sequence of images. 2D pose estimation is the basis of the majority of the pose estimation studies. According to the number of target animals in the scene that we are processing, we can divide the pose estimation problems into single animal pose estimation and multiple animal pose estimation. In terms of the input data to the models, these methods can be further classified as image-based vs. video-based, when a single image or a sequence of images is used, respectively.

3.1.1. Single Animal Pose Estimation

Individual pose estimation (single-task) means estimating for just one specific target in an image, whereas multiple pose estimation (multi-task) aims to estimate all target subjects in the image. The performance of animal pose estimation is primarily determined by its backbone as well as human pose estimation. We would like to briefly introduce two backbones widely used in animal pose estimation. The first one is Residual Neural Network (ResNet) (He et al., 2016) which was proposed in 2015. The residual block structure of ResNet enables the network to perform identity mapping between layers to avoid the exponential decay of the gradient correlation of the network due to the increasing of the number of layers. The simple and efficient structure makes it a viable backbone for many animal pose estimations (Li and Lee, 2021; Mu et al., 2020; Mathis et al., 2018). The second backbone, High-Resolution Network (HRNet) (Wang et al., 2020),is one of the best and most popular networks in both of human and animal pose estimation. As the depth of the network increases, HRNet continues to add parallel branches of low-resolution feature maps while keeping high-resolution feature maps. Accurate keypoint prediction can be achieved by fusing the semantic information in low-resolution feature maps and the precise feature which doesn't suffer information loss during down-sampling in high-resolution feature maps. (Yu et al., 2021a; Huang et al., 2021; Liu et al., 2022).

On the other hand, as we mentioned in Section 1, the performance of 2D pose estimation also depends heavily on handling the limited data. 2D single-task in the Table 1 can be divided into three types according to the solutions to data scarcity. The first is to use prior-aware synthetic data augmentation (Li and Lee, 2021; Mu et al., 2020). These works learn prior knowledge from massive synthetic animals datasets with labels and silhouettes. Then, unlabeled real data can be exploited by being assigned the pseudo-labels which are generated from the model trained with synthetic data. Another approach is to transfer learning from existing animals datasets or even humans (Cao et al., 2019). By defining the same labels based on the similar feature, a self-supervised joint training network can be constructed to refine the pseudo-labels. In addition, pre-training on large datasets such as ImageNet (Deng et al., 2009) has

Table 1: An overview and categorization of the animal pose estimation works from 2013 to 2021.

Year	First Authors	Animal Classes	Training Data Type	Research Direction	Highlights
2021	Russello et al. (2021)	Cow	bounding box, 2D annotations	2D pose	Based on LEAP (Pereira et al., 2019), T-LEAP improves the robustness to the varying background by increasing the depth of LEAP neural network.3D convolution is applied to spatial-temporal dimension to leverage the temproal information in videos and deal with occlusions.
2021	Mathis et al. (2021)	Horse	2D annotations	2D pose	The generalization ability for pose estimation and the robustness to image corruptions are analyzed on Mobilenetv2 (Sandler et al., 2019), EfficientNets (Tan and Le, 2020),ResNets (He et al., 2016).
2021	Joska et al. (2021)	Cheetah	2D annotations	2D pose and 3D pose	Provide a huge cheetahs dataset in the wild through using multi-view synchronized high-speed camera system and DeepLabCut for 2D annotation; 3D pose estimation is also produced by using three methods.
2021	Zhang et al. (2021)	Mouse	2D annotations	3D pose	Hierarchical von Mises-Fisher-Gaussian model that incorporates prior distributions of spatiotemporal constraints to make keypoint predictions based on Bayesian inference.
2021	Li and Lee (2021)	Horse, Tiger	2D annotations	2D pose	Multi-scale domain adaptation module (MDAM) to solve the domain shift between synthetic data and real data, combined with a pseudo label updating strategy to preserve annotations through the domain shift.
2020	Biggs et al. (2020)	Dog (120 breeds)	2D annotations, silhouettes, 3D priors	3D mesh (provide 3D pose and 2D pose)	End to end method for dog's 3D pose and shape estimation from single images based on SMAL; good generalization to new domain.
2020	Bala et al. (2020)	Rhesus macaque	2D annotations	3D pose	Multiview 3D reconstruction with augmented annotated data from 62 cameras, used to train a view invariant pose detector containing a deep neural network, that predicts 3D pose.
2020	Liu et al. (2020)	Mouse, Zebra, Monkey	2D annotations	3D pose	Video-based pose estimation architecture that combines a flexible base model called FlexibleBaseline to account for shape variety and an optical flow model to interpret video frames, to generate enhanced keypoint predictions.
2020	Zhang and Park (2020)	Monkey, Dog	2D annotations	3D pose	Semi-supervised model is trained with only less than 4% labeled data under cross-view supervision, temporal supervision and visibility supervision.
2020	Mu et al. (2020)	Quadruped Mammal	synthetic CAD model, pose segmentation, 2D and 3D joints	2D pose	Generate synthetic dataset from CAD animal models to mitigate the lack of labeled data; consistency-constrained semi-supervise learning is used to bridge the domain gap between synthetic and real data.
2019	Zuffi et al. (2019)	Zebra, horse	2D joints, Silhouettes, 3D priors	3D mesh (provide 3D pose and 2D pose)	Model is trained with synthetic dataset with real instance and synthesized pose, shape, texture, and background; Texture prediction is linked to 3D pose and shape through a shared feature space.
2019	Cao et al. (2019)	Quadruped Mammal	2D annotations, bounding box	2D pose	Find the shared features between human and animals and learn from their labeled data to estimate the pose of unseen categories; make full use of unlabeled data through progressive pseudo-label-based optimization.
2019	Graving et al. (2019a)	Grévy's zebra, Desert locust, Vinegar fly	2D annotations	2D pose	Animal pose estimation toolbox based on Stacked DenseNet deep learning model; The processing speed is 2 times faster than DeeplabCut through using Stack DenseNet and a fast GPU-based peak-detection method.
2019	Pereira et al. (2019)	Mouse	2D annotations	2D pose	Toolbox containing a graphical user interface for body-part labeling in images and a deep convolution neural network that produces probability distributions for the locations of each body part.
2018	Zuffi et al. (2018)	Quadruped Mammal	2D annotations, Silhouettes	3D mesh	Initial mesh which has been aligned to one image is optimized through being fitted to other images on multi-view; Model's texture is recovered from images by defining a UV map of texture coordinates.
2018	Biggs et al. (2018b)	Quadruped Mammal	2D annotations, Silhouettes	3D mesh (provide 3D pose & 2D pose)	2D joints are regressed from silhouettes by using multimodal heatmaps; 2D-to-3D correspondences are defined through optimal joint assignment and minimize the complex objective by using genetic algorithm (GA).
2018	Mathis et al. (2018)	Mouse	bounding box, 2D annotations	2D pose and 3D pose	The animal pose estimation toolbox based on DeeperCut feature detector architecture which can get human-level labeling accuracy by being trained only with 200 frames.
2017	Zuffi et al. (2017)	Quadruped Mammal	2D joints, Silhouettes	3D mesh	Model is trained with 3D scans of toy figurines; GLoSS model: shape deformations are performed in each part locally.
2016	Reinert et al. (2016)	Quadruped Mammal	2D joints, Silhouettes	3D mesh	The skeletal sketch is tracked through image sequences by using optical flow; Mesh is composed of aligned "generalized 3D cylinder" for each limb.
2015	Kanazawa et al. (2015)	Cat, Horse	2D joints, 3D Priors , silhouettes	3D mesh	Deform the mesh based on the local stiffness which value is various from skull to joints instead of ARAP.
2015	Ntouskos et al. (2015)	Giraffe	silhouettes	3D mesh	Animal is segmented into several components and modeled separately.
2013	Vicente and Agapito (2013)	Giraffe, Dolphins, Goose	Silhouettes	3D mesh	Template based 3D recovery method which uses silhouette based method to recovery the template mesh from single images.

also been shown to improve model accuracy and generalization (Mathis et al., 2018, 2021; Yu et al., 2021a). We would discuss these solutions in detail in the Section 6. Besides, Zhang and Park (2020) also reduces the need for labeled data by building a multi-camera systems for 2D animal pose estimation. Based on the supervision of spatio-temporal continuity of the keypoints movement and the characteristic that one pose in multi-view data must satisfy epipolar constraints (Hartley and Zisserman, 2003), they can train the end-to-end neural network with extensively unlabeled data and few real images. The result proves that the high precision pose estimation of both non-human and human subjects can be realized with highly limited labeled data. However, the multi-view camera system required by this method makes it difficult to be extended to animal pose estimation in the wild.

3.1.2. Multiple Animal Pose Estimation

Compared to the single animal pose estimation problem, multiple pose estimation is conducive to analyzing and understanding the social interaction between animals. Multiple pose estimation is not a simple composition of individual pose estimation. Generally speaking, the methods of multiple pose estimation are divided into top-down and bottom-up. In a top-down network (e.g. Stacked Hourglass (Newell et al., 2016) and HRNet (Wang et al., 2020)), the target detection will be performed firstly on the image to find all individuals, then individual pose estimation will be applied on each individual within the identified regions. In other words, top-down means transforming the multiple pose estimation into several individual pose estimation. Although top-down usually achieves high precision because each individual is processed separately, it is likely to be affected when animals are occluded by one another (Lauer et al., 2021). On the other hand, bottom-up networks (e.g. OpenPose (Cao et al., 2019; Cao et al., 2018), Higher-Resolution Networks (Cheng et al., 2019)), will find all the keypoints in the image first, and then assemble the points to each individual. Obviously, the main error comes from wrong assembly. Inspired by OpenPose (Cao et al., 2018), Lauer et al. (2021) purposes a novel multi-task architecture DLCRNet. In the encoder, DLCRNet which adopts the structure of HRNet, also builds a multi-fusion architecture to fuse the feature maps with various resolutions on parallel branches. In the decoder, in addition to the branch which predicts the score map of keypoints, a part affinity field (Cao et al., 2017), which can predict the location and orientation of limbs, is used to link the specific keypoints within one animal.

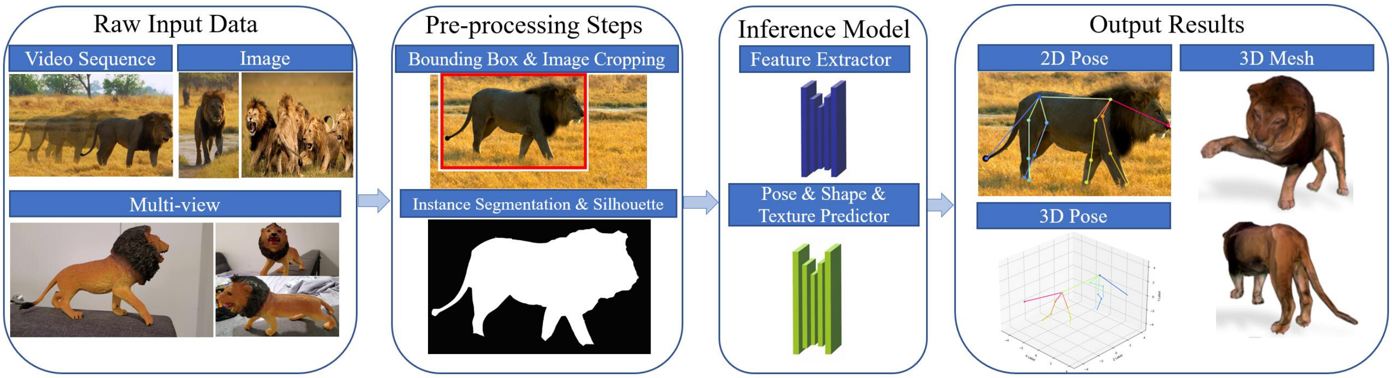

Fig. 2: An overview flow chart describing the generic workflow of existing animal pose or shape estimation methods. The input data of the networks can be video sequences or images with one or multiple animals in a monocular view or a multi-view. The raw data then can be pre-processed by being cropped or segmented, based on the requirement of each model. The model will learn the relevant representation by doing feature extraction and predict the pose, shape or texture based on the objective of the algorithm to output 2D/3D poses or 3D meshes.

3.1.3. Video-based Animal Pose Estimation

Video-based pose estimation requires a set of images consecutive in time and space as input for the inference model. It can produce more robust predictions by analyzing the temporal information and forcing temporal consistency in the image sequence. Russello et al. (2021) is a typical work of extracting temporal information from consecutive images to alleviate the environmental occlusion which often occurs in the wild. By adding one more dimension to the encoder and decoder of its original model LEAP (Pereira et al., 2019), the network can predict the pose of one image based on the feature maps extracted from its several adjacent images. The input of the new model is not a single image but a sequence of frames of length T while the output remains a single frame. Moreover, to cope with the increased data complexity, one additional 3D convolutional layer is added to the encoder and decoder of LEAP to extract more complex spatio-temporal features. The only problem is that they only test the model on the process of the cows walking from left to right.

The optical flow (Baker and Matthews, 2004) is another choice which is used in Zhang and Park (2020) to enforce the temporal consistency across an image stream. Optical flow is an excellent method to track the apparent motion of individual pixels on an image based on the color-immutability of a pixel in adjacent frames (Turaga et al., 2010). The continuously changing pose can be supervised by tracking the dense optical flow. In order to address the tracking drift caused by occlusion, Zhang and Park (2020) also wraps all of the keypoint as a whole to eliminate the argmax operation and apply the temporal supervision only on the frames with sufficient magnitude of the integral dense optical flow. Apart from these methods, in Table 1, most of the video-based pose estimation works are only based on the composition of pose estimation on single frame (Mathis et al., 2021, 2018; Biggs et al., 2018b). Compared to the pose estimation on single image, video-based pose analysis can not only be used to estimate poses, but also predict and analyze animal behavior and their group interactions by looking at dynamics of the pose data over time (Mathis et al., 2018; Bala et al., 2020).

3.2. 3D Animal Pose Estimation

3D animal pose estimation means predicting the joint positions of an animal in the 3D space from one or multiple images. Compared with the 2D pose, a more comprehensive pose description enables 3D pose estimation to be applied in wild aspects, such as animal ethology (Bala et al., 2020; Bauer et al., 2020), robotics (Joska et al., 2021; Peng et al., 2020). We divide the existing 3D works on animal pose estimation into multi-view vs. monocular (i.e. single view) 3D pose estimation and discuss them separately.

3.2.1. Multi-view 3D Pose Estimation

Multi-view pose estimation requires a camera system composed of multiple cameras to take several photos of animals synchronously. 2D pose estimation is performed on each perspective, and the 3D keypoints are calculated using the spatial relationship of the camera systems. Since the depth information of the keypoints can be easily calculated by using the calibration parameters of multiple cameras (Iskakov et al., 2019; Ummenhofer et al., 2017), the multi-view 3D pose estimation is still the most accurate method to obtain the 3D pose of animals.

In Bala et al. (2020), OpenMonkeyStudio, currently the best macaques deep learning-based markless motion capture system, is built to estimate the 3D pose (13 joints) of macaques. The system can observe the free movement of macaques from all directions through 62 synchronized high-definition cameras. The Convolutional Pose Machine (CPM) (Wei et al., 2016) is used to estimate the 2D pose in each image, from which the 3D pose is triangulated, and then the reconstructed 3D pose can be used to augment the 2D pose-labeled data. A large macaques dataset containing about 200k images is also proposed following. Similarly, Joska et al. (2021) builds a multi-view synchronized camera system with six GoPro cameras to collect data from 10 cheetahs in the wild. They utilize DeepLabCut (Mathis et al., 2018) as the 2D pose detector which is trained with AcinoSet dataset (Daniel et al., 2021). Then, they apply two more methods for multi-camera 3D pose estimation besides triangulation (Page, 2005), including Extended Kalman Filter (EKF) (Forsyth, 2002) and Full Trajectory Estimation (FTE) (Joska et al., 2021). Based on the result of the three baseline methods, they generate the first animal 3D pose in-the-wild benchmark.

It is true that we can obtain accurate 3D pose by using multi-camera system, but the difficulty of data collection is much higher than for monocular pose estimation. Most of the works have no choice but to collect data from animals in cages (Bala et al., 2020; Zhang and Park, 2020). It is challenging to set up and calibrate many cameras and collect data from animals in the wild. Therefore, the potential value of monocular pose estimation is immeasurable.

3.2.2. Monocular 3D Pose Estimation

Compared to the multi-view, monocular 3D pose estimation has far more unsolved challenges. Under the situation of 2D data scarcity, it even suffers from the more severe shortage of 3D pose-labeled data whose ground truth should be obtained through motion capture (MoCap) systems. Moreover, since one 2D pose can correspond to multiple possible 3D poses, elevating a 2D pose estimated in a single image to a 3D pose must face inherent ambiguity in depth. To accurately estimate the depth of keypoints, geometric prior knowledge of the target which can serve as a strong constraint needs to be learned first.

The SMAL model (Zuffi et al., 2017) which can generalize to many types of animals becomes the best option for providing the geometric prior. The monocular 3D pose estimation task can be simplified into fitting the pose, shape, and camera parameters of the parametric mesh. After reconstructing the animal's mesh from a single image, the 3D keypoint can be obtained from the coordinate of the vertex in the SMAL mesh. Due to the powerful generalization ability of the SMAL model, almost all the current monocular 3D pose estimation works are based on the SMAL model. In Zuffi et al. (2019), Zuffi et al. train an end-to-end neural network, SMAL with learned Shape and Texture (SMALST), using synthetic Grevy's zebra data. After that, Biggs et al. (2020) proposed an end-to-end neural network, Skinned Multi-Breed Linear Model for Dogs (SMBLD), suitable for a wide variety of dogs by enriching the shape parameters of the SMAL and learning more shape priors. We present the above works in detail in the 3D Animal Mesh Recovery.

Besides, Liu et al. (2021) also summarizes monocular human pose estimation methods from 2014 to 2021. Although there is no large 3D pose-labeled animal datasets like human do, such as Human3.6M (Ionescu et al., 2014; Catalin Ionescu, 2011), there are still many ideas which we can learn from. For example, Chen et al. (2019) proposed an unsupervised 2D-to-3D lifting method to eliminate the rely on 3D ground truth. This method first projects the lifted 3D poses from random views back to 2D images to supervise depth estimation using generated adversarial 2D poses. They then lift the randomly projected 2D pose to 3D and supervise the difference between its 2D projection from the original view and the original 2D pose input. This way, the model can be properly trained without any 3D annotations. Last but not least, we can also get synthetic 3D data and its ground truth from 3D animal meshes easily. A large number of 2D-3D pose pairs can be generated by projecting the 3D pose back to the 2D image plane, which can conducive to lifting 2D Pose to 3D (Chen and Ramanan, 2017).

3.3. 3D Animal Mesh Recovery

3D mesh reconstruction is a long-standing but still challenging task. At present, we can easily reconstruct the 3D mesh from the object's point cloud, which can be collected through RGBD cameras (Vyas et al., 2019) and 3D scanners (Duncan et al., 2017) or even be calculated from dozens of multi-view photos of objects (Agisoft, 2022). These methods are very convenient for rigid objects or cooperative humans, but they are hard to be conducted on non-cooperative, live and deformable animals. Therefore, recovering 3D mesh from single-view pictures is ideal for studying animals. Like pose estimation, 3D animal mesh recovery also plays a vital role in behavior studies of pets, wild animals and endangered species (Biggs et al., 2020; Zuffi et al., 2019; Youwang et al., 2021). This section will summarize the work of 3D animal mesh recovery from a single image, especially on the most popular method, model-based methods.

3.3.1. Model-based Methods

The core of the model-based approach is a parameterized deformable template mesh, whose parameters include mesh shape, pose, deformation, etc. The shape and pose parameters are estimated from a single image to compute the deformation of the template mesh, and the deformed mesh is constrained and optimized through comparing the intermediate estimations, such as silhouettes, 2D keypoint and camera parameters, against their ground truth. In the past, people would use rough meshes made by 3D animators as deformable templates, such as dolphins (Cashman and Fitzgibbon, 2012a), cats (Kanazawa et al., 2015), horse (Kanazawa et al., 2015), birds (Kanazawa et al., 2018a). However, the quality of model-based methods largely depends on the quality of the parameterized template, which requires many 3D scans to learn. Therefore, the model-based method was hard to conduct prior to the invention of the SMAL model (Zuffi et al., 2017).

Currently, most of the animal 3D mesh reconstruction works are based on the SMAL model which is the animal version of Skinned Multi-Person Linear model (SMPL) (Loper et al., 2015), because of its powerful functionality. The best human body models are learned from thousands of 3D scans of human bodies, but that is infeasible with many live animals. Thus, Zuffi et al. (2017) trained the model with the scans of a set of animal toys. They built a template mesh which is segmented into 33 parts and fit the template mesh to diverse poses and shapes by controlling the shape and pose parameters of each part. Regarding model deformation, they proposed global/local stitched shape model (GLoSS), which locally defines deformation of each part and stitches the parts by minimizing a stitching cost at the their interfaces. After the GLoSS model is fitted to all 3D scans, Zuffi et al. learn the parametric mean model of all animals by computing the shape difference's principal component (PCA). Finally, the parametric SMAL mesh can be reconstructed from a single image by fitting the shape, pose, and camera parameters to the labels and silhouette. Later, they proposed SMAL with Refinement (SMALR) Zuffi et al. (2018) that improves on the original SMAL. SMALR requires images of animals from multiple perspectives and 2D their silhouettes as its input data. The 2D silhouettes and 2D keypoints from different perspectives are used as constraints to reduce the ambiguity during 3D pose estimation. Then, the 2D projection is re-projected to the 3D mesh to optimize the generated model and make it fit the images better. In Zuffi et al. (2017) and Zuffi et al. (2018), the 2D label and silhouette are still required as ground truth to constrain the deformation while the end-to-end neural network SMALST (Zuffi et al., 2019) was purposed later. Since the model is only trained with the data of a certain animal, grevy's zebra, 3D pose, shape and even texture can be regressed from a single grevy's zebra image without any annotation and silhouette.

Besides, Biggs et al. (2020) purposed a specific SMAL model called SMBLD for dogs, by adding 6 new shape parameters to account for the variation of dogs. For the large variation in shape and appearance between dog breeds, they learn a detailed 3D prior from a large-scale dog dataset (Benjamin et al., 2020). Also, the domain gap between the manual designed shape prior and the real data is also alleviated through regularly updating the means and variances for each mixture component and per-image mixture weights based on the observed shapes in the training set with expectation maximization (EM). However, the texture cannot be regressed from SMBLD. Youwang et al. (2021) even realized a stable multi-task to estimate SMPL and SMAL model at the same time by unifying the label of human and animal in the same image. In addition to recovering the model from single RGB image, we can also consider using the depth information of a single RGBD image to implement an additional weak supervision to strengthen the 2D supervision (Cai et al., 2018; Chen et al., 2021).

Silhouette is an important constraint in model-based method (Cashman and Fitzgibbon, 2012a; Kanazawa et al., 2018b; Zuffi et al., 2017, 2018; Biggs et al., 2018b). It plays a role in fitting the generated model by comparing the generated mesh's projection on a certain plane to the manually created silhouettes. Although there are some large animal datasets for classification, detection, and instance segmentation, they only cover a small part of the world's animal species. In fact, researchers still need to generate the silhouettes themselves. One method is using deep learning models for semantic image segmentation, as shown in Deeplabv3+ (Chen et al., 2018b) and Mask-RCNN (Waleed, 2017). However, the performance of these models will depend on the trained networks. The emergence of synthetic datasets can be an effective solution of this problem Biggs et al. (2018b); Zuffi et al. (2017, 2018, 2019); Mu et al. (2020). The SMAL model and animal's CAD model can accurately give the silhouettes of a large number of animals with different poses and even the ground truth of their 2D and 3D keypoints.

3.3.2. Other Mesh Recovery Methods

Mesh Recovery from Template Mesh recovery from template, which can be called template-based method, is a simple alternative to the model-based approach, without training a parameterized template mesh (Vicente and Agapito, 2013; Malti et al., 2013). Both of them are the mesh deformation methods. In the template-based method, The relationship between the image and the template mesh is only established through the known corresponding points. The geometric constraint, such as As-Rigid-As-Possible (ARAP) (Sorkine and Alexa, 2007) and Isometric and Conformal (Malti et al., 2013), is used to constrain the deformation space instead of calculating deformation based estimated pose and shape parameters. Template-based methods mainly aim to reconstruct a bounded surface which are fully visible in the image, such as a paper or face. In order to apply the method on more challenging objects, Vicente and Agapito (2013) purpose a method to reconstruct the closed surface without boundaries. They obtained the template closed surface from the image through a contour-based monocular mesh recovery method (Oswald et al., 2012). The closed surface template is deformed to fit the input image according to the silhouette, region constraints, and point correspondences. Although they can reconstruct the basic shape and pose of the animal from a single image conveniently, the generated model lacks accuracy. However, after the appearance of the human model SMPL and general animal model SMAL, the accuracy and generalization of the model increases, and the cost of training a parameterized model decreases, which make model-based methods become mainstream gradually.

Mesh Recovery from Generalized Cylinders The 3D mesh can also be restored from a few simple geometric shapes,such as cuboids or generalized cylinders (Terzopoulos et al., 1988). Based on it, Gingold et al. (2009) and Reinert et al. (2016) purpose the pipeline to recover the mesh from the combination of the generalized cylinders with user-defined sketch drawn in the video. It allows the user to draw strokes along a body part of a creature (torso, limbs, tail, etc.) on a sequence of sparse key frames in the video. Optical flow (Pérez et al., 2013) is then used to automatically track and expand the strokes to all frames and instance is segmented on the basis of strokes through the cost-volume filtering (Hosni et al., 2012). Then, cylindrical fitting is performed on each part of the body according to the strokes and segmentation. This method does not require any 3D prior and can be widely used for a variety of animals, even giraffes and elephants. However, due to the lack of depth information in the video (they do not use animal model like SMAL as a template), the 3D error of the generated model could be relatively higher than that of model-based method which has strong geometric prior.

4. Existing Animal Pose Datasets

The development of new animal pose estimation models is more than often accompanied by release of new animal pose datasets. Unlike the human pose estimation, the physical diversity among animal species makes it difficult for researchers to use the datasets composed of different subjects. Although several works aim to overcome the cross-domain gaps between different species Yu et al. (2021a); Cao et al. (2019), their performances are obviously worse than those which use specific animals for training. Also, due to the lack of a widely accepted standard, the types of annotations vary from one dataset to another. In this section, we introduce exiting easily accessible and open-source animal datasets, as tabulated in Table 2.

Table 2: Summary of publicly-available animal pose datasets. Majority of them are annotated image/video datasets, with only a few unannotated image and 3D mesh datasets. The missing elements in each dataset is highlighted in red.

Year	Dataset	Animal Classes	Size	Data Type	Annotation	First Released/ Introduced
2021	AP-10K Dataset (Yu et al., 2021a)	23 animal families and 54 species	10K labeled images, 50K unlabeled images	2D images	17 Landmarks (2D pose)	(Yu et al., 2021b)
2021	Horse-10 Dataset (Rogers et al., 2021)	Horses	8.1K images	2D images	22 Landmarks	Mathis et al. (2021)
2021	AcinoSet (Daniel et al., 2021)	Cheetahs	119K frames	Multi-view video, 2D images	20 Landmarks (2D pose), 3D Pose Prediction	Joska et al. (2021)
2020	StanfordExtra (Benjamin et al., 2020)	Dogs	12K images	2D images	20 landmarks (2D pose)	Biggs et al. (2020)
2020	Synthetic Animal Dataset (Mu et al., 2019)	Synthetic Hound, Tiger, Horse, Sheep, Elephant	50K	2D images	18 Landmarks (2D pose), Segmentation	Mu et al. (2020)
2019	ATRW dataset (Jianguo et al., 2019)	Amur tigers	8k images	2D images	15 Landmarks (2D pose)	Li et al. (2019)
2019	Synthetic Grevy's Zebra Dataset(Silvia et al., 2019)	Synthetic grevy's zebra	13K images	2D images	28 Landmarks (2D pose)	Zuffi et al. (2019)
2019	Animal Pose (Jinkun et al., 2019)	Dogs, Cats, Cows, Horses, Sheep, the other 7 categories	710 MB	2D images	20 Landmarks (2D pose), Bounding boxes	Cao et al. (2019)
2018	Animals-10 (Alessio, 2018a)	Dog, Cat, Horse, Sheep, Cow, Elephant	586 MB	2D images	No	Alessio (2018b)
2018	BADJA (Biggs et al., 2018a)	Bears, Camels, Cows, Dogs, Horses, Tigers, Cats	2.8GB	2D frames from video	20 Landmarks (2D pose), Segmentation	Biggs et al. (2018b)
2017	Digital Life 3D (Duncan et al., 2017)	Rhino, Lizard, Dolphin	N/A	3D mesh (with texture)	No	Duncan et al. (2017)
2016	TigDog (Pero et al., 2016)	Dogs, Horses, Tigers	7.5 GB	2D video	19 Landmarks (2D pose)	Del Pero et al. (2015b) Del Pero et al. (2017)
2014	COCO (Lin et al., 2014a)	Cat, Dog, Horse, Sheep, Cow,Elephant, Bear, Giraffe, Zebra	18GB	2D images	Bounding Box, Segmentation, Class label	Lin et al. (2014b)
2012	Poselets (Bourdev, 2012)	Dogs, Cats, horses, Sheep, Cows	1.6GB	2D images	16 Landmarks (2D pose), Bounding box	Bourdev (2012)
2011	Stanford Dogs (Aditya et al., 2011)	Dogs(120 breeds, 20K)	2.8GB	2D images	Class labels, Bounding box	Khosla et al. (2011) Deng et al. (2009)
2009	Non-rigid world (Alexander et al., 2009)	Cats, Dogs, Wolves, Horses,Lions, Gorillas	24MB	3D mesh (without texture)	No	Bronstein et al. (2006) Bronstein et al. (2007)

4.1. Image-based Animal Pose Datasets

The image-based animal pose dataset consists of individual images that are not related to each other. That is to say, these images usually contain a large number of animal instance with different textures and shapes, and an assortment of diverse backgrounds, lighting, camera angles, and occlusions. Therefore, it is constructive for training the generalization ability, texture robustness, and occlusion robustness of the model with image-based dataset.

AP-10K (Yu et al., 2021a) is the latest and largest labeled dataset for general animal pose estimation. This dataset, which was just released in 2021, consists of 10,000 labeled images and 50,000 unlabeled images of 23 animal families and 54 species, covering a large number of common animal family , such as Canidae, Felidae , Bovidae, Equidae, Ursidae, and Cercopithecidae. The labels are consistent with the COCO dataset labels (Lin et al., 2014b), including 2 eyes, 1 nose, 1 neck, 1 root of tail, and 3 joints in each limb. 13 well-trained annotators are recruited to annotate all the keypoints and three rounds of cross-checking and correction are then carried out to ensure the high quality of the annotation. Since the labels are provided following the COCO dataset style, we can do transfer learning easily between the human dataset and animal dataset (Cao et al., 2019). In addition, the rich animal species and diverse animal posture, backgrounds, and occlusions will greatly facilitate the generalization ability of animal pose estimation models intra-family and even inter-family. However, due to the considerable differences in physical characteristics between species and between animals and humans, it still remains to be seen whether the label entirely consistent with the human dataset (Lin et al., 2014a) will meet the requirements for animal research.

Poselets (Bourdev, 2012) is a publicly available dataset built by Lubomir Bourdev in 2009. It is mainly composed of the annotated human instance from PASCAL VOC 2007 to 2011. It also includes the keypoints and bounding box annotations for all animal instance from PASCAL 2011. In the dataset, there are 1266 cats, 642 cows, 1571 dogs, 760 horses and 878 sheep. The keypoints include 2 eyes, 2 ears, 1 nose, 4 paws, 4 elbows, 1 throat, 1 withers, and 1 tailbase. All of the annotations were manually labeled by using the H3D annotation tools (Lubomir and Jitendra, 2011).

Animals-10 (Alessio, 2018b) is a large animal image dataset, which is collected by Corrado Alessio from google images. It contains 6 categories of mammal: dog (4863 images), cat (1668 images), horse (2623 images), sheep (1820 images), cow (1866 images), and elephant (1446 images). Unfortunately, this dataset does not have any pose labels. However, the large collection of categorized images can be conducive to the establishment of general animal pose dataset, like Animal Pose dataset (Jinkun et al., 2019).

Animal Pose (Cao et al., 2019) is a collection of Poselets dataset whose annotations were extended by Jinkun Cao et al to 20 keypoints (original 16 keypoints plus 4 knees). In addition, they also extended the size of this dataset through annotating more images from the Animals-10 dataset. This dataset has two subsets. Subset 1 only includes the five categories in Poselets and it provides more than 6000 pose-labeled instances in 4608 images. The annotation in subset 1 has already been aligned to the format of MS COCO. In subset 2, there are only bounding box annotations for other 7 animal categories, such as deer, monkey, hippo, bear and rhino. Due to its large amount of data, specialized labels, rich and representative species, this dataset is often used in the training of models with cross-domain adaptation (Cao et al., 2019; Mathis et al., 2021) and the evaluation of model generalization ability (Mathis et al., 2021; Li and Lee, 2021).

Stanford Dogs (Khosla et al., 2011; Deng et al., 2009) is a large dog dataset released in 2011 by Aditya Khosla (Khosla et al., 2011). This dataset contains 20,580 images of 120 dog breeds and each class has more than 150 images. There are class labels and bounding box annotations for each image. These images were generated from ImageNet (Deng et al., 2009) and verified by Amazon Mechanical Turk (MTurk). Although there is no pose label data, the comprehensiveness of the canine family's images contained in this dataset make it a good research material for object segmentation and pose estimation (Biggs et al., 2020).

StanfordExtra (Biggs et al., 2020) is a new large-scale dog dataset with 2D keypoints and instance segmentation annotations released in 2020. The authors extracted 8,476 images (which are suitable for keypoint annotation and 3D reconstruction) from the Stanford Dogs dataset, and annotate 20 keypoints and binary silhouette for each image through MTurk. The keypoints include 2 ear-base, 2 ear-tip, 1 nose, 1 jaw, 1 tail-base, 1

Table 3: Pose estimation accuracy (in PCK@0.05 metric) of a few models applied on the horse and tiger portion of the TigDog dataset. The first and second rows indicate the results of ResNet-50 (He et al., 2016) trained only with real labeled data or synthetic labeled data. CC-SSL (Mu et al., 2020) and MDAM (Li and Lee, 2021) models both represent the models trained with labeled synthetic data and unlabeled real data.

Method	**Training Data**	**PCK@0.05 Accuracy on Horse Data**								**PCK@0.05 Accuracy on Tiger Data**							
		Eye	Chin	Shoulder	Hip	Elbow	Knee	Hoove	Average	Eye	Chin	Shoulder	Hip	Elbow	Knee	Hoove	Average
ResNet-50	Real	79.04	89.71	71.38	**91.78**	**82.85**	80.80	72.76	78.98	96.77	**93.68**	**65.90**	**94.99**	**67.64**	**80.25**	**81.72**	**81.99**
ResNet-50	Synthetic	46.08	53.86	20.46	32.53	20.20	24.20	17.45	25.33	23.45	27.88	14.26	52.99	17.32	16.27	19.29	21.17
CC-SSL	Synthetic&Real	84.60	90.26	69.69	85.89	68.58	68.73	61.33	70.77	96.75	90.46	44.84	77.61	55.82	42.85	64.55	64.14
MDAM	Synthetic&Real	**91.05**	**93.37**	**77.35**	80.67	73.63	**81.83**	**73.67**	**79.50**	**97.01**	91.18	46.63	78.08	50.86	61.54	70.84	67.67

tail-tip and 3 joints in each limb. This dataset is one of the few datasets that takes into account the peculiarities of animal studies and thus adds labels such as ear and tail that are not common in human datasets. Due to the inheritance of the comprehensive canine breeds of the stanford dogs dataset, researchers can train the model with strong shape and appearance robustness.

4.2. Video-based Animal Pose Datasets

Video-based animal pose datasets are often obtained by shooting a few animal instances continuously in order to study a particular species. Because of the fixed camera angle, the small number of objects, and the monotonous background, it is hard to train a model with strong robustness on a video-based dataset. However, consecutive shots can make it easier for researchers to obtain more significant amounts of data and various poses. And, in addition to pose estimation and object segmentation, video data can also be used for animal behavior analysis.

BADJA (Biggs et al., 2018b) is a video-based animal dataset of 2D pose annotation released in 2018 (Biggs et al., 2018b). This dataset is composed of two parts. In the first part, there are 7 video sequences which are from the DAVIS video segmentation dataset Pont-Tuset et al. (2017). The other 4 video sequences in the second part were collected individually and segmented by using Adobe's UltraKey tool Adobe (2018). There are 8 animal categories in total: bear, camel, cat, dog, horse, cow, impala, and tiger. Each video sequence is no more than 100 frames and each was annotated every 5 frames with 20 keypoints and visibility indicators. The keypoints include 2 ears, 1 nose, 1 jaw, 3 for tail (from base to tip), 1 neck, 4 knees, 4 ankles, and 4 toes.

TigDog (Del Pero et al., 2015b, 2017) is a large video-based animal dataset which was released with Del Pero et al. (2015b) 2015 and Del Pero et al. (2017) 2016. All of the video sequences were annotated with behavior labels, keypoints, and the instance segmentation. Tigdog includes 3 animal categories: dog, horse and tiger. 19 keypoints are annotated in all 16,000 frames of the horse class, and in 17,000 of the tiger class.The annotations include: eyes (2), neck (1), chin (1), hooves (4), hips (4) and knees (4). For tigers: eyes (2), neck (1), chin (1), ankles (4), feet (4) and knees (4). The ability of this dataset to train models with strong generalization and robustness has been verified in several works (Mu et al., 2020; Li and Lee, 2021).

AcinoSet (Joska et al., 2021) is a cheetah video dataset, which contains 119,490 frames taken by multi-view synchronized high speed cameras in the wild, released in 2021 (Joska et al., 2021). Its 7,588 frames are manually annotated as ground truth. Besides, all 119K frames have the predicted 2D keypoints, which are estimated through DeepLabCut (Mathis et al., 2018). Furthermore, 3D pose estimation was also conducted in this work; 3D trajectories (generated through Full Trajectory Estimation (Joska et al., 2021)) and human-checked 3D keypoints ground truth are provided.

ATRW (Li et al., 2019) is the Amur Tiger Re-identification in the Wild dataset. More than 8,000 Amur tiger footage and 92 instances are collected from around 10 zoos in China. The bounding box, 15 keypoints and individual identity are manually labeled in this dataset. It will be greatly conducive to the conservation, information preservation of endangered species

Horse-10 Dataset (Rogers et al., 2021) is composed of more than 8,000 frames extracted from video footage captured with GoPro camera. It includes 30 different thoroughbred horses and most of them have more than 200 annotated poses. There are 22 landmarks on each horse: nose (1), eye (1), each leg (3), shoulder (1), mid-shoulder (1), hip (1), girth (1), elbow (1), wither (1), stifle (1), ischium (1). All the frames were annotated through DeepLabCut 2.0 toolbox (Nath et al., 2019). The diversity of horse breeds is helpful for the research on generalization ability of the pose estimation model.

4.3. 3D Animal Model Datasets

3D animal model datasets have always served as the training data for 3D reconstruction works Zuffi et al. (2017, 2018); Mu et al. (2020). Their production often requires sophisticated 3D scanners or 3D animators.

Non-rigid World (Bronstein et al., 2006, 2007) is a high-resolution non-rigid 3D shapes dataset. This dataset has 7 kinds of animal models, including cat (11 poses), dog (9 poses), wolf (3 poses), horse (17 poses), lion (15 poses), and gorilla (21 poses). All of the models are manually designed. In Kanazawa et al. (2015), the cat and horse models were used as the template 3D surface in the 3D mesh deformation.

Digital Life 3D (Duncan et al., 2017) aims to create precise and high-resolution 3D models of animals on the earth for animal conservation and education, providing both paid and free 3D models. The paid models are made by 3D animators and even contain fantasy creatures. Meanwhile, the free models are generated through scanning the real animals. University of Massachusetts at Amherst built a photography system consisting of 20 cameras and placed the animal, such as a turtle, in the center. The coarse models are generated by using photogrammetry software RealityCapture. 3D artists would use Blender to improve the coarse models and get the final model of the animal.

4.4. Synthetic Animal Pose Datasets

As we mentioned above, a common problem in animal pose or shape estimation is the scarcity of data. Generally it is difficult to obtain large amounts of real data with the diverse pose, shape, and camera angle, which is required to train an accurate model using deep learning, as manually labeling data is expensive and time-consuming. Therefore, we tend to recommend a promising substitute, synthetic data, for real data in this section.

Synthetic Grevy's Zebra Dataset (Silvia et al., 2019) is the training set generated in Zuffi et al. (2019). In Zuffi et al. (2019), they first reconstructed 10 zebra models from 57 images by using 3D mesh recovery pipeline SMALR (Zuffi et al., 2018). Next, they randomly changed the pose, texture and shape of these 10 models to enrich the model set. The pose of each zebra model can be changed by adjusting 3D Rodrigues vectors which describe the model's pose. Further, noise is randomly added to the brightness, hue, saturation levels, shape and camera focus to vary the appearance, shape and size of zebras. 12,850 images with different backgrounds are rendered by using neural 3D mesh renderer (Kato et al., 2018). In addition to the basic labels of the COCO dataset, it also includes ear, tail, cheek and mouth in the 28 landmarks, which makes it one of the richest datasets in terms of labels. Since these labels are automatically generated by projecting the corresponding points on the 3D mesh into 2D images, they can be changed and added label at any point on the mesh.

Synthetic Animal Dataset (Mu et al., 2019) is composed of 50K synthetic images generated in (Mu et al., 2020). Unlike Zuffi et al. (2019), CAD animal models are used as the template and the images are rendered with Unreal Engine 4. There are 5 species in the dataset, including hound, tiger, horse, sheep, and elephant. All of the labels provided are consistent with the TigDog dataset except the neck. Compared with models trained with real data, training with synthetic animal dataset has better domain generalization performance in multiple visual domains (Mu et al., 2020). Table 3 shows that both Consistency-Constrained Semi-Supervised Learning (CC-SSL) (Mu et al., 2020) and Multi-scale Domain Adaptation Module (MDAM) (Li and Lee, 2021) models which are trained with labeled synthetic data and unlabeled real data have comparable performance with the supervised model, ResNet-50 (He et al., 2016), trained only with the real data and even surpass it for some parts of the horse in TigDOg dataset (Pero et al., 2016).

5. Performance Evaluation of Animal Pose Estimation Models

Animal pose estimation problem involves many different species and datasets. Here, we further summarize the type of results, test datasets, evaluation protocols, and main experimental results in Table 4 according to works listed in Table 1. Due to the diversity of the evaluation metrics in each work, it is difficult to know and compare the state-of-the-art on animal pose estimation. Therefore, we introduce several evaluation metrics which are widely used at present: root mean square error (RMSE), percentage of correct keypoints (PCK), intersection over union (IoU), and mean average precision (mAP). Finally, we compare and visualize the results of several methods with common evaluation metrics.

5.1. Evaluation Metrics

PCK was first introduced by Yang and Ramanan (2012) as a better substitute for the probability of a correct pose (PCP) to evaluate the joint localization accuracy. It shows the percentage of the predicted keypoints which fall within a normalized threshold of the ground truth. The PCK of the i^{th} keypoint in N targets is calculated as the following equation.

$$PCK_i = \frac{\sum_n \delta(\frac{d_{n_i}}{s_n} < \alpha)}{\sum_n 1}, \tag{1}$$

where, n means the the n^{th} target; d_{n_i} is the the Euclidean distance between the i^{th} predicted keypoint and its ground truth of the n^{th} target; α is a constant parameter which controls the relative correctness of the evaluation and s_n is a normalized scalar of the n^{th} target. α and s are various among different works. For instance, in the FLIC dataset (Sapp and Taskar, 2013),the Euclidean distance from the left shoulder to the right hip or the Euclidean distance from the right button to the left hip is defined as the normalized scalar while the MPII dataset (Andriluka et al., 2014b) employs Euclidean distance of the diagonal of the bounding box of the person's head as the scale factor. However, because of the various shape of animal, the length of head or torso is difficult to be widely adopted as a fair normalizer unless there is only a single species in this work (Mathis et al., 2021). For the reasons above, the square root of the area of the data can be considered a fair scale factor and be widespread in animal pose estimation (Li and Lee, 2021; Mu et al., 2020; Zuffi et al., 2019; Mathis et al., 2021). The threshold is distributed between 0.05 and 0.3. And PCK@0.05 is usually considered a high precision regime.

mAP, which is based on Object Keypoint Similarity (OKS) became another commonly used metric for pose estimation after the appearance of COCO keypoint challenge (Lin et al., 2022). OKS is a metric to evaluate the similarity between the predicted keypoints with the ground truth and Average Precision (AP) is employed to see the performance of the prediction on the dataset by giving a threshold which is compared against OKS. mAP is the mean of multiple AP which have different threshold. OKS and AP is defined as:

$$OKS = \frac{\sum_i exp(\frac{-d_i^2}{2s^2k_i^2})\delta(v_i > 0)}{\sum_i \delta(v_i > 0)}, \tag{2}$$

$$AP = \frac{\sum_n \delta(OKS_n > T)}{\sum_n 1}, \tag{3}$$

where, d_i is the the Euclidean distance between the i^{th} predicted keypoint and its ground truth; s is the scale factor of this target, whose value is equal to the square root of the area of the target's bounding box; k_i is the special constant, which controls fall off of the i^{th} keypoint and COCO dataset; (Lin et al., 2014a) provides special constants for the 17 different labels; v_i shows the visibility of the annotated keypoint which can be 0, 1, 2 for unlabeled, labeled but occluded and labeled and visible respectively; OKS_n is the OKS of the n^{th} target; T is the given threshold.

PCK aims to judge whether the prediction is correct or not for each key point separately, while mAP pays more attention to the overall evaluation of all key points of a target and makes the evaluation more accurate by giving each key point a different scalar. If the predicted dataset has the same label as the

Table 4: Performance evaluation of the existing animal pose estimation models. In the type of result, KP means keypoint. The missing elements in each dataset is highlighted in red.

Year	First Authors	Type of result	Test Dataset	Evaluation Protocol	Result
2021	Russello et al. (2021)	2D KP	CoWalk-10 and CoWalk-30 dataset (Russello et al., 2021)	PCKh@0.2 (%)	CoWalk-30: No-occlusion=99.0, 3 occlusions=88.4; CoWalk-10: known cow=93.8, unknown cow=87.6
2021	Mathis et al. (2021)	2D KP	Horse-10 and Horse-C (Rogers et al., 2021)	PCKh@0.3 (%)	Efficientnet-B6: known horse=99.9; unknown horse=88.4
2021	Joska et al. (2021)	3D KP	AcinoSet (Daniel et al. (2021))	RMSE,SEM,NRSME (pixels)	Run: RMSE=28.24; Dive: RMSE=76.35
2021	Zhang et al. (2021)	3D KP	Multi-view video (Zhang et al., 2021)	Mean Position Error (mm)	Mouse=5.77
2021	Li and Lee (2021)	2D KP	TigDog (Pero et al. (2016)), VisDA2019 (Kate et al. (2019))	PCK@0.05 (%)	Horse=79.5; Tiger=67.76
2020	Biggs et al. (2020)	3D mesh, 3D KP	StanfordExtra (Benjamin et al. (2020)), Animal pose (Jinkun et al. (2019))	IoU (%), PCK@0.15 (%),	IOU= 74.2; PCK= 78.8
2020	Bala et al. (2020)	3D KP	OpenMonkeyPose dataset	head location error (m)	mean error=0.0714; standard deviation=0.0234
2020	Liu et al. (2020)	2D KP	N/A	aPCK error (%) (Liu et al., 2020)	Mouse≈ 10; $Monkey \approx 0.4$
2020	Zhang and Park (2020)	2D KP	Multi-view images on human and animal (Zhang and Park, 2020)	AUC on PCKh (%)	Human=95.1; Dog=94.8; Monkey=92.2
2020	Mu et al. (2020)	2D KP	TigDog (Pero et al. (2016)), VisDA2019 (Kate et al. (2019))	PCK@0.05 (%)	Horse=82.43; Tiger=84
2019	Zuffi et al. (2019)	3D mesh, 3D KP	200 annotated zebra image (Zuffi et al. (2019))	PCK@0.05 (%), PCK@0.1 (%), IOU (%)	PCK@0.05=62.3; PCK@0.1=81.2; IOU=42.2
2019	Cao et al. (2019)	2D KP	Animal Pose (Jinkun et al. (2019)), COCO Dataset (Lin et al. (2014b))	mAP (%)	mAP=65.7
2019	Graving et al. (2019a)	2D KP	Fly,locust, Grevy's zebra dataset(Graving et al. (2019b))	RMSE (pixels)	Zebra=1.85
2019	Pereira et al. (2019)	2D KP	Fly dataset (Pereira et al., 2018)	RMSE (pixels)	Fly=1.63 per 47μm
2018	Zuffi et al. (2018)	3D Mesh	Frames from GreenScreenAnimal or YouTube	3 views of each mesh	No Quantitative Evaluation
2018	Biggs et al. (2018b)	3D Mesh, 2D KP	BADJA	PCK@0.2 (%)	Felidae=95; Canidae=87.4; Equidae=89.8; Bovidae=95; Hippopotamidae=93.9
2018	Mathis et al. (2018)	2D KP	Fly and mouse hand dataset (Mathis et al. (2018))	RMSE (pixels)	Fly=4.17 ± 0.32; Mouse hand= 5.21 ± 0.28
2017	Zuffi et al. (2017)	3D mesh, 3D KP	TigDog, images from internet	Mean normalized distance error	Horse: 0.068(5 images), Rhino:0.069(5 images)
2016	Reinert et al. (2016)	3D Mesh	N/A	IoU (%)	IOU 3D: Horse=63.53; Camel=58.31; IoU 2D: Horse=85.85; Camel=83.95
2015	Kanazawa et al. (2015)	3D Mesh	Non-rigid world (Alexander et al. (2009))	1 view mesh	No Quantitative Evaluation
2015	Ntouskos et al. (2015)	3D Mesh	Image from Flickr, model from Warehouse-SketchUp	Mean normalized Hausdorff distance error (Aspert et al., 2002)	Cat=0.012; Dog=0.012; Cow=0.03; Sheep=0.04; Hippo=0.013; Giraffe=0.018
2013	Vicente and Agapito (2013)	3d Mesh	Dolphin Dataset (Cashman and Fitzgibbon (2012b))	2 views of each mesh	No Quantitative Evaluation

COCO dataset, then mAP can be calculated using COCO-api (Lin, 2020; Cao et al., 2019). However, if the dataset labels are different from the COCO dataset, the scalar for the labels according to the dataset must be calculated. Due to the uncertainty of animal labels, mAP is not as universal as PCK in animal pose estimation.

RMSE is a metric to calculate the Euclidean distance error in pixels between model-generated annotation with the manual annotation (ground truth). RMSE can show the annotation error at a specific body part, averaged over the whole body, or averaged over a image sequence. It is always used by the works with such high accuracy that PCK can hardly evaluate the performance of the model. Take DeepLabCut (Mathis et al., 2018) as example, the error can be reduced to 5 pixels when 100 frames are used to train the model. RMSE is now widely accepted as the main evaluation metric by the "out-of-the-box" animal pose estimation toolboxes and the work based on them, as it is used to evaluate DeepLabCut, DeepPoseKit and LEAP (Mathis et al., 2018; Graving et al., 2019a; Pereira et al., 2019).

IoU is the overlap rate between the generated candidate bound (CB) and the ground truth bound (GB), that is, the ratio of their intersection and union. The ideal situation is complete overlap, then the ratio would be 1. This is a metric for detection and it can also be used to evaluate the generated 3D mesh through calculating the IoU of the projection of generated 3D mesh on certain plane.

$$IoU = \frac{\text{area of } [CB \cap GB]}{\text{area of } [CB \cup GB]}. \tag{4}$$

5.2. Results Comparison and Visualization

In this section, we compare three cases of animal pose estimation algorithms and visualize their results using common evaluation metrics.

Case #1: In this case, we mainly focus on three recent animal pose estimation toolboxes. By comparing their RMSE after each training iteration as well as their training time, we can make a preliminary evaluation on the performance of these three toolboxes. In Graving et al. (2019a), the DeepPoseKit, DeepLabCut, and LEAP toolboxes were each given 9 iterations of model training (with optimization to the model after each iteration) and the RMSE was recorded at the end of each iteration. Across the 9 iterations, DeepPoseKit had the lowest RMSE (Fig. 3). Training time is another important metric used to compare animal pose estimation toolboxes and as Fig. 4 shows the training time required to reach the minimum RMSE was lowest for DeepPoseKit. As such, it is accurate to say that DeepPoseKit is the best performing toolbox among the models we examined.

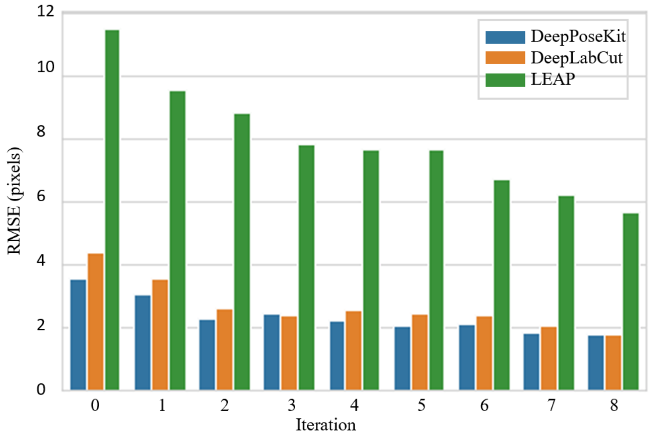

Fig. 3: RMSE performance comparison of DeepLabCut, DeepPoseKit, and LEAP, after being applied on the same fruit fly, locust and zebra 2D pose-labeled datasets (Graving et al., 2019a)

.

Case #2: In this case, we mainly visualize the results from six works which use PCK as their evaluation protocol in Fig. 5. As mentioned above, the value of PCK obtained from the same job will increase when its relative correctness becomes larger. Therefore, it is difficult to assess which work has better performance by comparing some works using different correctness PCK as protocol. Generally speaking, a work will have better performance if the relative correctness is lower and the accuracy is higher.

Case #3: In this case, we show several results on 3D mesh recovery. These works only evaluated the shape estimation, instead of texture. In the Fig. 6, we can see that the sketching 3D mesh recovery method utilized by Reinert et al. (2016) achieves a good performance on horse and camel data, especially in 2D IoU which reached the state of the art. This is because the sketching method does not have any prior information. The generalized cylinder is used to construct the animal's body, and the cylinder is fitted according to the segmentation of the animal. On the other hand, some model-based methods, like SMAL, is constrained by the template mesh. Therefore, the sketching method can achieve high 2D IoU, but less details on the mesh. Also, the 3D IoU value would be relatively low due to the lack of depth information and 3D prior. Then for dog's mesh recovery, Biggs et al. (2020) accomplishes the highest IoU, 74.2%, which surpasses the 70.5% produced by SMAL pipeline (Zuffi et al., 2018). Besides, SMALST pipeline (Zuffi et al., 2019) reached the state-of-the-art on the 3D mesh recovery of Grevy's zebra. However, due to the lack of real data from Grevy's zebra, the network is trained with synthetic data. The IoU produced by SMALST on Grevy's zebra is 42.2%.

6. Discussion on Existing Challenges

So far, we have described different techniques of animal pose estimation that have been developed over the past few years and categorized them based on the algorithms they use to accomplish this task. We have learned that the data/label scarcity is comparable to a dark cloud shrouding an uncultivated field. Due to the lack of sufficient annotated data for training, some models used for human pose estimation find it difficult to achieve the same performance in animal pose estimation. Although more large-scale labeled datasets for general animal pose estimation are being released (Yu et al., 2021a), the species included is a drop in the ocean for all animal on the earth. Therefore, at the end of the article, we further the discussion on this issue and analyze plausible methods available to overcome these challenges.

6.1. Data/Label Scarcity

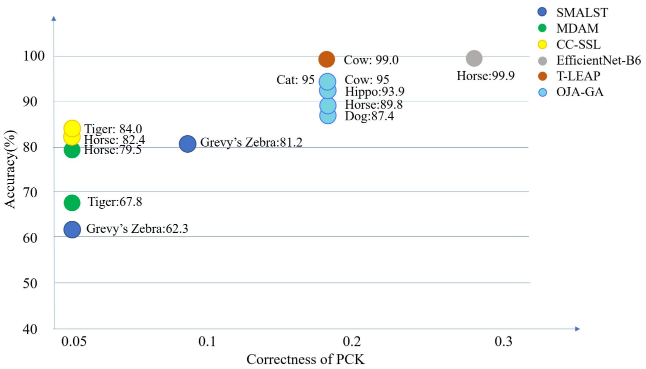

Fig. 5: Comparison of the performance of six animal pose estimation approaches. SMALST is the method in Zuffi et al. (2019), MDAM represents Li and Lee (2021), CC-SSL represents Mu et al. (2020), EfficientNet-B6 is used in Mathis et al. (2021), T-LEAP is introduced in Russello et al. (2021) and OJA-GA denotes Biggs et al. (2018b).

Data scarcity can be subdivided into two types. The most common one is label scarcity. Due to the huge difference in texture and shape between species, pose or shape estimation research on certain animal requires a corresponding labeled dataset. Although some researchers are able to create specialized datasets for their own research (Joska et al., 2021; Li et al., 2019), unfortunately, the time-consuming and laborious labeling task still hinders many studies. The other type is data scarcity. It is difficult even to collect enough unlabeled data. They are commonly seen in studies of endangered species in the wild (Zuffi et al., 2019; Li et al., 2019).

6.2. Prominent Solution

One promising solution is to use massive synthetic datasets, which are automatically annotated, to make up a deficiency of real data. In addition, the domain gap between synthetic and real animal data is easier to manage than the gap between other domains such as humans and animals (Li and Lee, 2021). In Zuffi et al. (2019), due to the lack of real Grevy's zebra data in the wild, they chose to train the model with a large synthetic dataset generated by a 3D animal models recovery pipeline (Zuffi et al., 2018) based on few real data, from which the 2D/3D ground truth and silhouettes can be produced automatically. This greatly reduces the time required to annotate pose

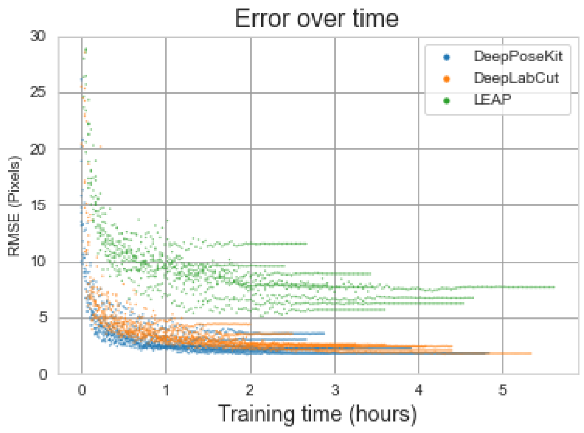

Fig. 4: Comparison of the performance of DeepLabCut, DeepPoseKit, and LEAP by runtime.

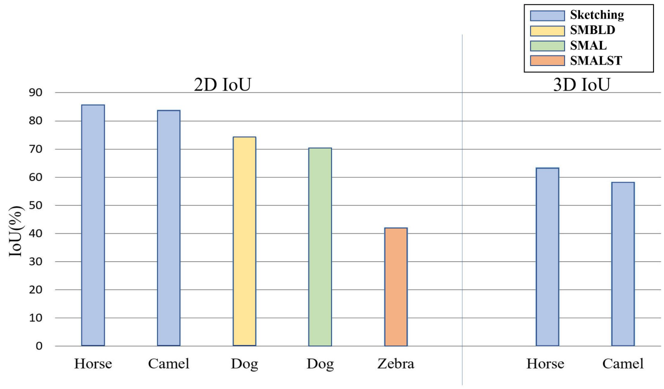

Fig. 6: Comparison of the performance of four 3D mesh recovery approaches for four animals. Sketching is the method in Reinert et al. (2016), SMBLD represents Biggs et al. (2020), SMAL represents Zuffi et al. (2017) and SMALST is Zuffi et al. (2019). On the left side, the 3D mesh is evaluated on 2D IoU; On the right side, the mesh is evaluated on 3D IoU.

data. Similarly, in order to solve the lack of labeled data, Mu et al. (2020) we can generate a large general animal synthetic dataset with CAD models and Unreal Engine and designed a model (Consistency-Constrained Semi-Supervised Learning) which can be trained with a mix of labeled/unlabeled real data and synthetic data. Coarse labels are assigned to unlabeled images based on prior knowledge learned from synthetic data and the coarse labels are continuously optimized during joint training on real and synthetic data. The accuracy of joint training on labeled real data and synthetic data can even go beyond the model trained only with real data while joint training on unlabeled real data and synthetic data can achieve a close performance. However, due to the domain gap between synthetic data and real data, the pseudo labels generated by the model trained with the synthetic data will be noisy and difficult to filter out based on confidence score. Mu et al. (2020), Li and Lee (2021) further improved the model trained with mixed datasets by alleviating the domain shift in the data using a domain classifier to learn domain invariant features. We showed their results in Table 3. In addition to using the above two method, generating a synthetic dataset from animal toys is an another convenient and affordable way. There are many existing works such as (Liu and Ostadabbas, 2018; Vyas et al., 2021), which have already built pipelines to generate 3D Synthetic model from mannequins or toys and rig them in software (Blender). By rendering the synthetic model with varying pose, shape, lighting, camera, and background, we can easily obtain large amount of animal image in different perspective and environment without any real data.

Another approach is transferring learning from an existing real dataset. The intention of transfer learning is to save the cost of manually labeling by transfer the feature domain of the unlabeled data (target domain data) to existing labeled data (source domain data). A model suitable for the target domain can be trained without much labeled target domain data. This is based on learning prior knowledge from numerous labeled human data due to the similarity between humans and quadruped. In Cao et al. (2019), the author propose a cross-domain adaptation to enforce the existing massive labeled human data and labeled/unlabeled animal data to share the same feature space. By unifying the label format of humans and the different kinds of animals, the pseudo labels can be estimated for the unlabeled animal data based on the model trained with human and labeled animal data. Then, the joint and alternating training with human, labeled/unlabeled animals data can be conducted to update the pseudo labels. Through mixing the human and several categories datasets, the data requirement of one certain category can be reduced. Similarly, Youwang et al. (2021) also finds the morphological similarity between the human and quadruped. They categorized physical corresponding body parts exist among human and quadruped, such as arms and legs, and defined them as sub-keypoints. The reconstruction loss of sub-keypoints is used as a bridge to realize a stable multi-task with both SMPL (Loper et al., 2015) and SMAL (Zuffi et al., 2017). Transfer learning between similar quadrupeds also becomes remarkable after the issue of multi-species datasets for animal pose estimation, like AP-10K (Yu et al., 2021a) and Animal Pose (Jinkun et al., 2019).

Training an accurate model with only a small amount of labeled data is also an attractive direction. In recent years, animal pose estimation toolboxes, such as DeepLabCut, DeepPoseKit and LEAP (Graving et al., 2019a; Pereira et al., 2019; Mathis et al., 2018), which are based on the state-of-the-art pose estimation model such as DeeperCut (Insafutdinov et al., 2016), DenseNet (Pereira et al., 2019) were purposed one by one. DeepLabCut (Mathis et al., 2018) demonstrates that pretraining on ImageNet (Deng et al., 2009) enables the model to have good object segmentation and keypoint regression capabilities before training on the dataset. By using the model pretrained on ImageNet, 200 annotated frames are enough to produce a model with ability of achieving human-level labeling accuracy (Nath et al., 2019). In addition, (Mathis et al., 2021; Yu et al., 2021a) show that pretraining on ImageNet or human dataset will increase the accuracy of intra-species estimation and the generalization of inter-species. Like using synthetic data, another benefit of using these toolboxes is customization. Users can freely define the labels and the toolbox will track the labels automatically in the video. This will allow researchers to design more appropriate labels for the animals they study without being limited by the existing large labeled datasets.

7. Conclusion

In this paper, we accumulated and categorized the research on animal pose or shape estimation from computer vision perspective mainly published between 2013 to 2021. We classified these techniques into three categories based on their final objective, including 2D pose estimation, 3D pose estimation, and 3D mesh recovery. For each category, we discussed the most prominent approaches and different methods used to solve the task at hand. We also looked at several animal pose and mesh datasets available publicly while exploring the use of synthetic data generation for animals that are difficult to capture in the wild. Then, we introduced the state-of-the-art animal pose estimation performance and compared these techniques using popular evaluation metrics used in the computer vision literature. Finally, we listed some of the major difficulties that researchers face when trying to solve animal pose estimation problem efficiently and explored certain promising solutions. Although, we are excited by the recent progress that has been made in this domain, there is still a lot of room for further development.

References

Aditya, K., Nityananda, J., Bangpeng, Y., Fei-Fei, L., 2011. Stanford dogs dataset. `http://vision.stanford.edu/aditya86/ImageNetDogs/`.

Adobe, 2018. Adobe systems inc.: Creating a green screen key using ultra key. `https://helpx.adobe.com/premiere-pro/user-guide.html`.

Agisoft, L., 2022. Photoscan. `https://www.agisoft.com/`.

Alessio, C., 2018a. Animals-10 dataset. `https://www.kaggle.com/alessiocorrado99/animals10`.

Alessio, C., 2018b. Deep learning for image recognition. `https://www.dropbox.com/s/j5xdb2hmkp900gz/corrado-tesina-en.pdf?dl=0`.

Alexander, B., Michael, B., Ron, K., 2009. Project tosca:tools for non-rigid shape comparison and analysis. `http://tosca.cs.technion.ac.il/book/resources_data.html`.

Anderson, T.L., Donath, M., 1990. Animal behavior as a paradigm for developing robot autonomy. Robotics and autonomous systems 6, 145–168.

Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B., 2014a. 2d human pose estimation: New benchmark and state of the art analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B., 2014b. 2d human pose estimation: New benchmark and state of the art analysis. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) .

Aspert, N., Santa-Cruz, D., Ebrahimi, T., 2002. Mesh: measuring errors between surfaces using the hausdorff distance, in: Proceedings. IEEE International Conference on Multimedia and Expo, pp. 705–708 vol.1. doi:10.1109/ICME.2002.1035879.

Baker, S., Matthews, I., 2004. Lucas-kanade 20 years on: A unifying framework. International journal of computer vision 56, 221–255.

Bala, P.C., Eisenreich, B.R., Yoo, S.B.M., Hayden, B.Y., Park, H.S., Zimmermann, J., 2020. Automated markerless pose estimation in freely moving macaques with openmonkeystudio. Nature communications 11, 1–12.

Bauer, S., Klaassen, M., 2013. Mechanistic models of animal migration behaviour–their diversity, structure and use. Journal of Animal Ecology 82, 498–508.

Bauer, U., Poppinga, S., Müller, U.K., 2020. Mechanical ecology—taking biomechanics to the field. Integrative and Comparative Biology 60, 820–828.

Benjamin, B., Oliver, B., James, C., 2020. 3d animal reconstruction with expectation maximization in the loop. https://sites.google.com/view/wldo/home.

Biggs, B., Boyne, O., Charles, J., Fitzgibbon, A., Cipolla, R., 2020. Who left the dogs out? 3d animal reconstruction with expectation maximization in the loop, in: European Conference on Computer Vision, Springer. pp. 195–211.

Biggs, B., Roddick, T., Fitzgibbon, A., Cipolla, R., 2018a. Badja. https://github.com/benjiebob/BADJA.

Biggs, B., Roddick, T., Fitzgibbon, A., Cipolla, R., 2018b. Creatures great and smal: Recovering the shape and motion of animals from video, in: Asian Conference on Computer Vision, Springer. pp. 3–19.

Blender, . https://www.blender.org/.

Bourdev, L., 2012. Dataset of keypoints and foreground annotations for all categories of pascal 2011. https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/shape/poselets/.

Bronstein, A.M., Bronstein, M.M., Kimmel, R., 2006. Efficient computation of isometry-invariant distances between surfaces. SIAM Journal on Scientific Computing 28, 1812–1836.

Bronstein, A.M., Bronstein, M.M., Kimmel, R., 2007. Calculus of nonrigid surfaces for geometry and texture manipulation. IEEE Transactions on Visualization and Computer Graphics 13, 902–913.

Butail, S., Abaid, N., Macrì, S., Porfiri, M., 2015. Fish–robot interactions: robot fish in animal behavioral studies, in: Robot fish. Springer, pp. 359–377.

Cai, Y., Ge, L., Cai, J., Yuan, J., 2018. Weakly-supervised 3d hand pose estimation from monocular rgb images, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 666–682.

Cao, J., Tang, H., Fang, H.S., Shen, X., Lu, C., Tai, Y.W., 2019. Cross-domain adaptation for animal pose estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9498–9507.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y., 2018. Openpose: real-time multi-person 2d pose estimation using part affinity fields. arXiv preprint arXiv:1812.08008 .

Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A., 2019. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence .

Cao, Z., Simon, T., Wei, S.E., Sheikh, Y., 2017. Realtime multi-person 2d pose estimation using part affinity fields, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7291–7299.

Cashman, T.J., Fitzgibbon, A.W., 2012a. What shape are dolphins? building 3d morphable models from 2d images. IEEE transactions on pattern analysis and machine intelligence 35, 232–244.

Cashman, T.J., Fitzgibbon, A.W., 2012b. What shape are dolphins? building 3d morphable models from 2d images. IEEE transactions on pattern analysis and machine intelligence 35, 232–244.

Catalin Ionescu, Fuxin Li, C.S., 2011. Latent structured models for human pose estimation, in: International Conference on Computer Vision.

Chen, C.H., Ramanan, D., 2017. 3d human pose estimation= 2d pose estimation+ matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7035–7043.

Chen, C.H., Tyagi, A., Agrawal, A., Drover, D., Stojanov, S., Rehg, J.M., 2019. Unsupervised 3d pose estimation with geometric self-supervision, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5714–5724.

Chen, K., Gabriel, P., Alasfour, A., Gong, C., Doyle, W.K., Devinsky, O., Friedman, D., Dugan, P., Melloni, L., Thesen, T., Gonda, D., Sattar, S., Wang, S., Gilja, V., 2018a. Patient-specific pose estimation in clinical environments. IEEE Journal of Translational Engineering in Health and Medicine 6, 1–11.

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018b. Encoder-decoder with atrous separable convolution for semantic image segmentation, in: ECCV.

Chen, S., Yang, R.R., 2020. Pose trainer: Correcting exercise posture using pose estimation. arXiv:2006.11718.

Chen, Y., Tu, Z., Kang, D., Bao, L., Zhang, Y., Zhe, X., Chen, R., Yuan, J., 2021. Model-based 3d hand reconstruction via self-supervised learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10451–10460.

Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T.S., Zhang, L., 2019. Bottom-up higher-resolution networks for multi-person pose estimation. arXiv preprint arXiv:1908.10357 7.

Daniel, J., Liam, C., Naoya, M., Ricardo, J., Fred, N., Alexander, M., Mathis, M.W., Amir, P., 2021. Acinoset: A 3d pose estimation dataset and baseline models for cheetahs in the wild. https://github.com/African-Robotics-Unit/AcinoSet.

Del Pero, L., Ricco, S., Sukthankar, R., Ferrari, V., 2015a. Articulated motion discovery using pairs of trajectories, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Del Pero, L., Ricco, S., Sukthankar, R., Ferrari, V., 2015b. Articulated motion discovery using pairs of trajectories, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2151–2160.

Del Pero, L., Ricco, S., Sukthankar, R., Ferrari, V., 2017. Behavior discovery and alignment of articulated object classes from unstructured video. International journal of computer vision 121, 303–325.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee. pp. 248–255.

Duncan, J.I., Andy, D., Neil, H., Paul, L., Angela, R., 2017. Digitallife3d:3d models. http://digitallife3d.org/3d-model.

Flickr, . Flickr. https://www.flickr.com/.

Forsyth, D.A., 2002. J. ponce computer vision–a modern approach.

Gingold, Y., Igarashi, T., Zorin, D., 2009. Structured annotations for 2d-to-3d modeling, in: ACM SIGGRAPH Asia 2009 papers, pp. 1–9.

Graving, J.M., Chae, D., Naik, H., Li, L., Koger, B., Costelloe, B.R., Couzin, I.D., 2019a. Deepposekit, a software toolkit for fast and robust animal pose estimation using deep learning. Elife 8, e47994.

Graving, J.M., Chae, D., Naik, H., Li, L., Koger, B., Costelloe, B.R., Couzin, I.D., 2019b. Fast and robust animal pose estimation. bioRxiv , 620245.

Güler, R.A., Neverova, N., Kokkinos, I., 2018. Densepose: Dense human pose estimation in the wild, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7297–7306.

Hao-Shu Fang, Shuqin Xie, Y.W.T., Cewu Lu, C.L., 2017. Alphapose. https://github.com/MVIG-SJTU/AlphaPose.

Hartley, R., Zisserman, A., 2003. Multiple view geometry in computer vision. Cambridge university press.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Hosni, A., Rhemann, C., Bleyer, M., Rother, C., Gelautz, M., 2012. Fast cost-volume filtering for visual correspondence and beyond. IEEE Transactions on Pattern Analysis and Machine Intelligence 35, 504–511.

Huang, X., Fu, N., Liu, S., Ostadabbas, S., 2021. Invariant representation learning for infant pose estimation with small data, in: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), IEEE. pp. 1–8.

Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B., 2016. Deepercut: A deeper, stronger, and faster multi-person pose estimation model, in: European Conference on Computer Vision, Springer. pp. 34–50.

Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C., 2014. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence 36, 1325–1339.

Iskakov, K., Burkov, E., Lempitsky, V., Malkov, Y., 2019. Learnable triangulation of human pose, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

Jianguo, L., Weiyao, L., Tang, H., Greg, M., Joachim, D., 2019. Iccv 2019 workshop & challenge on computer vision for wildlife conservation (cvwc). `https://cvwc2019.github.io/challenge.html`.

Jinkun, C., Hongyang, T., Hao-Shu, F., Xiaoyong, S., Cewu, L., Yu-Wing, T., 2019. Animal pose dataset. `https://sites.google.com/view/animal-pose/`.

Joska, D., Clark, L., Muramatsu, N., Jericevich, R., Nicolls, F., Mathis, A., Mathis, M.W., Patel, A., 2021. Acinoset: A 3d pose estimation dataset and baseline models for cheetahs in the wild. arXiv preprint arXiv:2103.13282 .

Kanazawa, A., Kovalsky, S., Basri, R., Jacobs, D., 2015. Learning 3d articulation and deformation using 2d images. CoRR .

Kanazawa, A., Tulsiani, S., Efros, A.A., Malik, J., 2018a. Learning category-specific mesh reconstruction from image collections, in: Proceedings of the European Conference on Computer Vision (ECCV).

Kanazawa, A., Tulsiani, S., Efros, A.A., Malik, J., 2018b. Learning category-specific mesh reconstruction from image collections, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 371–386.

Kate, S., Xingchao, P., Ben, U., Kuniaki, S., Ping, H., 2019. Visual domain adaptation challenge (visda-2019). `https://ai.bu.edu/visda-2019/`.

Kato, H., Ushiku, Y., Harada, T., 2018. Neural 3d mesh renderer, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3907–3916.

Khosla, A., Jayadevaprakash, N., Yao, B., Li, F.F., 2011. Novel dataset for fine-grained image categorization: Stanford dogs, in: Proc. CVPR workshop on fine-grained visual categorization (FGVC), Citeseer.

Lauer, J., Zhou, M., Ye, S., Menegas, W., Nath, T., Rahman, M.M., Di Santo, V., Soberanes, D., Feng, G., Murthy, V.N., et al., 2021. Multi-animal pose estimation and tracking with deeplabcut. bioRxiv .

Li, C., Lee, G.H., 2021. From synthetic to real: Unsupervised domain adaptation for animal pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1482–1491.

Li, S., Li, J., Tang, H., Qian, R., Lin, W., 2019. Atrw: a benchmark for amur tiger re-identification in the wild. arXiv preprint arXiv:1906.05586 .

Li, X., Zhang, J., Yin, M., 2014. Animal migration optimization: an optimization algorithm inspired by animal migration behavior. Neural Computing and Applications 24, 1867–1877.

Lin, T.Y., 2020. cocoapi. `https://github.com/cocodataset/cocoapi`.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014a. Coco dataset. `https://cocodataset.org/#download`.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014b. Microsoft coco: Common objects in context, in: European conference on computer vision, Springer. pp. 740–755.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2022. Coco keypoint challenge. `https://cocodataset.org/#home`.

Liu, S., Huang, X., Fu, N., Li, C., Su, Z., Ostadabbas, S., 2022. Simultaneously-collected multimodal lying pose dataset: Enabling in-bed human pose monitoring. IEEE Transactions on Pattern Analysis and Machine Intelligence .

Liu, S., Ostadabbas, S., 2018. A semi-supervised data augmentation approach using 3d graphical engines. `arXiv:1808.02595`.

Liu, W., Bao, Q., Sun, Y., Mei, T., 2021. Recent advances in monocular 2d and 3d human pose estimation: A deep learning perspective. arXiv preprint arXiv:2104.11536 .

Liu, X., Yu, S.y., Flierman, N., Loyola, S., Kamermans, M., Hoogland, T.M., De Zeeuw, C.I., 2020. Optiflex: video-based animal pose estimation using deep learning enhanced by optical flow. BioRxiv .

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J., 2015. SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) 34, 248:1–248:16.

Lubomir, B., Jitendra, M., 2011. The human annotation tool. `https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/shape/hat/`.

Malti, A., Hartley, R., Bartoli, A., Kim, J.H., 2013. Monocular template-based 3d reconstruction of extensible surfaces with local linear elasticity, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Mathis, A., Biasi, T., Schneider, S., Yuksekgonul, M., Rogers, B., Bethge, M., Mathis, M.W., 2021. Pretraining boosts out-of-domain robustness for pose estimation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 1859–1868.

Mathis, A., Mamidanna, P., Cury, K.M., Abe, T., Murthy, V.N., Mathis, M.W., Bethge, M., 2018. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. Nature neuroscience 21, 1281–1289.

Mu, J., Qiu, W., Hager, G.D., Yuille, A.L., 2019. Synthetic animal dataset. `https://github.com/JitengMu/Learning-from-Synthetic-Animals`.

Mu, J., Qiu, W., Hager, G.D., Yuille, A.L., 2020. Learning from synthetic animals, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12386–12395.

Nath, T., Mathis, A., Chen, A.C., Patel, A., Bethge, M., Mathis, M.W., 2019. Using deeplabcut for 3d markerless pose estimation across species and behaviors. Nature protocols 14, 2152–2176.

Neverova, N., Novotny, D., Khalidov, V., Szafraniec, M., Labatut, P., Vedaldi, A., 2020. Continuous surface embeddings. `arXiv:2011.12438`.

Newell, A., Yang, K., Deng, J., 2016. Stacked hourglass networks for human pose estimation, in: European conference on computer vision, Springer. pp. 483–499.

Nguyen, N.H., Phan, T.D.T., Lee, G.S., Kim, S.H., Yang, H.J., 2020. Gesture recognition based on 3d human pose estimation and body part segmentation for rgb data input. Applied Sciences 10.

Ntouskos, V., Sanzari, M., Cafaro, B., Nardi, F., Natola, F., Pirri, F., Ruiz, M., 2015. Component-wise modeling of articulated objects, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2327–2335.

Obdrzalek, S., Kurillo, G., Han, J., Abresch, T., Bajcsy, R., 2012. Real-time human pose detection and tracking for tele-rehabilitation in virtual reality, Studies in Health Technology and Informatics. pp. 320–324.

Oswald, M.R., Töppe, E., Cremers, D., 2012. Fast and globally optimal single view reconstruction of curved objects, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 534–541.

Ovsjanikov, M., Ben-Chen, M., Solomon, J., Butscher, A., Guibas, L., 2012. Functional maps: a flexible representation of maps between shapes. ACM Transactions on Graphics (TOG) 31, 1–11.

Page, G., 2005. Multiple view geometry in computer vision, by richard hartley and andrew zisserman, cup, cambridge, uk, 2003, vi+ 560 pp., isbn 0-521-54051-8.(paperback£ 44.95). Robotica 23, 271–271.

Peng, X.B., Coumans, E., Zhang, T., Lee, T.W., Tan, J., Levine, S., 2020. Learning agile robotic locomotion skills by imitating animals. arXiv preprint arXiv:2004.00784 .

Pereira, T.D., Aldarondo, D.E., Willmore, L., Kislin, M., Wang, S.S.H., Murthy, M., Shaevitz, J.W., 2018. Dataset:fast animal pose estimation using deep neural networks. `http://arks.princeton.edu/ark:/88435/dsp01pz50gz79z`.

Pereira, T.D., Aldarondo, D.E., Willmore, L., Kislin, M., Wang, S.S.H., Murthy, M., Shaevitz, J.W., 2019. Fast animal pose estimation using deep neural networks. Nature methods 16, 117–125.

Pérez, J.S., Meinhardt-Llopis, E., Facciolo, G., 2013. Tv-l1 optical flow estimation. Image Processing On Line 2013, 137–150.

Pero, L.D., Susanna, R., Rahul, S., Vittorio, F., 2016. Tigdog dataset. `http://calvin-vision.net/datasets/tigdog/`.

Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L., 2017. The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 .

RealityCapture, . https://www.capturingreality.com/.

Reinert, B., Ritschel, T., Seidel, H.P., 2016. Animated 3d creatures from single-view video by skeletal sketching., in: Graphics Interface, pp. 133–141.

Ricardo, T., 2022. Motion capture library. `http://www.mocapclub.com/Pages/Library.htm`.

Rogers, B., Mathis, A., Mathis, M.W., 2021. Horse-10: an animal pose estimation benchmark for out-of-domain robustness. `http://www.mackenziemathislab.org/horse10`.

Russello, H., van der Tol, R., Kootstra, G., 2021. T-leap: occlusion-robust pose estimation of walking cows using temporal information. arXiv preprint arXiv:2104.08029 .

Sanakoyeu, A., Khalidov, V., McCarthy, M.S., Vedaldi, A., Neverova, N., 2020. Transferring dense pose to proximal animal classes, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C., 2019. Mobilenetv2: Inverted residuals and linear bottlenecks. `arXiv:1801.04381`.

Sapp, B., Taskar, B., 2013. Modec: Multimodal decomposable models for human pose estimation, in: In Proc. CVPR.

Silvia, Z., Angjoo, K., Tanya, B.W., Michael, J.B., 2019. Training data of smalst. `https://github.com/silviazuffi/smalst`.

Sorkine, O., Alexa, M., 2007. As-rigid-as-possible surface modeling, in: Symposium on Geometry processing, pp. 109–116.

Sung, K.K., Poggio, T., 1998. Example-based learning for view-based human face detection. IEEE Transactions on pattern analysis and machine intelligence 20, 39–51.

Tan, M., Le, Q.V., 2020. Efficientnet: Rethinking model scaling for convolutional neural networks. `arXiv:1905.11946`.

Terzopoulos, D., Witkin, A., Kass, M., 1988. Symmetry-seeking models and 3d object reconstruction. International Journal of Computer Vision 1, 211–221.

Tsiktsiris, D., Dimitriou, N., Lalas, A., Dasygenis, M., Votis, K., Tzovaras, D., 2020. Real-time abnormal event detection for enhanced security in autonomous shuttles mobility infrastructures. Sensors 20.

Turaga, P., Chellappa, R., Veeraraghavan, A., 2010. Advances in video-based human activity analysis: challenges and approaches, in: Advances in Computers. Elsevier. volume 80, pp. 237–290.

Ummenhofer, B., Zhou, H., Uhrig, J., Mayer, N., Ilg, E., Dosovitskiy, A., Brox, T., 2017. Demon: Depth and motion network for learning monocular stereo, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5038–5047.

Vicente, S., Agapito, L., 2013. Balloon shapes: Reconstructing and deforming objects with volume from images, in: 2013 International Conference on 3D Vision-3DV 2013, IEEE. pp. 223–230.

Vyas, K., Jiang, L., Liu, S., Ostadabbas, S., 2021. An efficient 3d synthetic model generation pipeline for human pose data augmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1542–1552.

Vyas, K., Ma, R., Rezaei, B., Liu, S., Neubauer, M., Ploetz, T., Oberleitner, R., Ostadabbas, S., 2019. Recognition of atypical behavior in autism diagnosis from video using pose estimation over time, in: 2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP), IEEE. pp. 1–6.

Waleed, A., 2017. Mask r-cnn for object detection and instance segmentation on keras and tensorflow. `https://github.com/matterport/Mask_RCNN`.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al., 2020. Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence .

Warehouse-SketchUp, D., . 3d, warehouse-sketchup. `https://3dwarehouse.sketchup.com/`.

Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y., 2016. Convolutional pose machines, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 4724–4732.

Yang, G., Huang, T.S., 1994. Human face detection in a complex background. Pattern recognition 27, 53–63.

Yang, Y., Ramanan, D., 2012. Articulated human detection with flexible mixtures of parts. IEEE transactions on pattern analysis and machine intelligence 35, 2878–2890.

Youwang, K., Ji-Yeon, K., Joo, K., Oh, T.H., 2021. Unified 3d mesh recovery of humans and animals by learning animal exercise. arXiv preprint arXiv:2111.02450 .

Yu, H., Xu, Y., Zeng, J., Zhao, W., Guan, Z., Tao, D., 2021a. Ap-10k: A benchmark for animal pose estimation in the wild. `https://github.com/AlexTheBad/AP-10K`.

Yu, H., Xu, Y., Zhang, J., Zhao, W., Guan, Z., Tao, D., 2021b. Ap-10k: A benchmark for animal pose estimation in the wild. arXiv preprint arXiv:2108.12617 .

Zhang, L., Dunn, T., Marshall, J., Olveczky, B., Linderman, S., 2021. Animal pose estimation from video data with a hierarchical von mises-fisher-gaussian model, in: International Conference on Artificial Intelligence and Statistics, PMLR. pp. 2800–2808.

Zhang, Y., Park, H.S., 2020. Multiview supervision by registration, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 420–428.

Zuffi, S., Kanazawa, A., Berger-Wolf, T., Black, M.J., 2019. 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, pp. 5359–5368.

Zuffi, S., Kanazawa, A., Black, M.J., 2018. 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, pp. 3955–3963.

Zuffi, S., Kanazawa, A., Jacobs, D.W., Black, M.J., 2017. 3d menagerie: Modeling the 3d shape and pose of animals, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6365–6373.