symmetry

MDPI

Article

A Large-Scale Mouse Pose Dataset for Mouse Pose Estimation

Jun Sun [1,†], Jing Wu [1,†], Xianghui Liao [1,†], Sijia Wang [1,†] and Mantao Wang [2,*]

[1] College of Information Engineering, Sichuan Agricultural University, Ya'an 625000, China; 2019319014@stu.sicau.edu.cn (J.S.); 201902236@stu.sicau.edu.cn (J.W.); 201902210@stu.sicau.edu.cn (X.L.); 201902197@stu.sicau.edu.cn (S.W.)
[2] Sichuan Key Laboratory of Agricultural Information Engineering, Ya'an 625000, China
* Correspondence: wangmantao@sicau.edu.cn
† These authors contributed equally to this work.

Abstract: Mouse pose estimations have important applications in the fields of animal behavior research, biomedicine, and animal conservation studies. Accurate and efficient mouse pose estimations using computer vision are necessary. Although methods for mouse pose estimations have developed, bottlenecks still exist. One of the most prominent problems is the lack of uniform and standardized training datasets. Here, we resolve this difficulty by introducing the mouse pose dataset. Our mouse pose dataset contains 40,000 frames of RGB images and large-scale 2D ground-truth motion images. All the images were captured from interacting lab mice through a stable single viewpoint, including 5 distinct species and 20 mice in total. Moreover, to improve the annotation efficiency, five keypoints of mice are creatively proposed, in which one keypoint is at the center and the other two pairs of keypoints are symmetric. Then, we created simple, yet effective software that works for annotating images. It is another important link to establish a benchmark model for 2D mouse pose estimations. We employed modified object detections and pose estimation algorithms to achieve precise, effective, and robust performances. As the first large and standardized mouse pose dataset, our proposed mouse pose dataset will help advance research on animal pose estimations and assist in application areas related to animal experiments.

Keywords: mouse pose estimation; dataset; deep learning; computer vision

Citation: Sun, J.; Wu, J.; Liao, X.; Wang, S.; Wang, M. A Large-Scale Mouse Pose Dataset for Mouse Pose Estimation. *Symmetry* **2022**, *14*, 875. https://doi.org/10.3390/sym14050875

Academic Editor: Pecchinenda Anna

Received: 21 March 2022
Accepted: 22 April 2022
Published: 25 April 2022

1. Introduction

Benefiting from the advancement of deep learning networks and the improvement of sensor camera technologies, pose estimations have dramatically developed in the computer vision community during recent years. Research on pose estimations is not limited to humans and hands, but also extends to animal pose studies. Animal pose estimations are a key step in animal behavior research. Related animal research has confronted a set of increasing demands in the neuroscience [1], genetics [2], pharmacology [3], and psychology [4] domains. Traditional analyses of animal poses rely on manual recognitions and analyses of videos. This does not meet the needs of current research. Therefore, researchers in various disciplines have come to rely on computer vision systems for precise and detailed estimations of the pose.

With the rapid prosperity and maturity of pose estimation technologies of humans [5–7], animal pose estimations have been introduced in recent years. As a more challenging task, animal pose estimations have been drawing substantial attention. Existing research on pose estimations of various animal species include mice [8], cattle [9], birds [10], pigs [11], chimpanzees [12], fruit flies [13], etc. Among these, mice are a mammalian species that is frequently used in bioscientific research. They have the features of a small size, fast growth, and low price and are easy to breed and use [14], and as such, they are widely used in various research fields. Therefore, it is necessary to study mouse pose estimations based on computer vision. It is well known that pose estimation technologies have matured for

human pose estimation applications. Diverse and precise systems have been proposed in human pose estimations, as well as behavior analyses [15–20]. However, due to different the physiological characteristics between humans and mice, the same methods cannot be migrated to mouse pose estimations directly. Specifically, the mouse is highly deformable, and its limbs are normally sheltered by its body. Therefore, it is a difficult task to make accurate, fast, and robust measurements of mice behaviors.

Thus far, there exists a range of algorithms, frameworks, and approaches on mouse pose estimations [21–24]. However, they are hindered either by diverse and possibly inconsistent principles or by unstandardized image data captured through different equipment. In detail, regarding the aspect of image data, they have been generated by retrieving from camera sensors [2,25] or by using existing publicly available datasets [26]. The quality of such training data is inconsistent. Hence, the need for a large and uniform dataset for estimating full mouse poses has emerged.

In this paper, we introduce a real-world novel and large-scale mouse pose dataset. The dataset was captured from continuous color video and ground-truth 2D behaviors among interacting mice. Profiting from 10 pairs of mice raised in a stable laboratory environment, we collected recorded top-view videos and extracted abundant frames. Various improvements were also made in both the number and the quality of mouse poses. Our mouse pose dataset can assist to advance the state-of-the-art mouse pose estimations and provide a wider range of possibilities for future research.

Under normal circumstances, the limbs of mice are obscured by their bodies. This phenomenon makes precise annotations a difficult problem. To address this problem, we creatively define five locations of keypoints: the mouth, the left ear, the right ear, the neck, and the tail root. Among these, the keypoint neck is located in the center of an image, while the other two pairs of keypoints are symmetric. This symmetric feature makes the keypoints conspicuous and simple to operate, as well as observe. In the 40,000 RGB images of the mouse pose dataset, accurate annotations were well labeled by us on the locations of the keypoints of the mice. Each picture shows the location of a mouse in detail: its bounding box, the X and Y image coordinates of its five joint positions. Meanwhile, diversity is completely demonstrated here. Various postures of multiple mice at different times profoundly expanded the profusion of our dataset, such as upright, climbing, feeding, etc.

We also designed a hardware device to collect the videos of the mice. The hardware device is equipped with a camera for videos' acquisition and an LED lamp for balance in illumination. Additionally, annotating data is an essential step for the training of neural networks and machine learning. The work of fast and accurate data annotation is a non-negligible long-lasting bottleneck of various applications in these fields. Despite the availability of software for annotating human datasets, they are not suitable to be used for mice. Underlying this fact, we developed specific software for annotating our mouse dataset. The software not only relatively alleviates the work of humans in this time-consuming and tedious task, but it is also easy to reproduce. It has potential wide applications in related work. For completeness, we present a baseline for mouse pose estimations based on the previous work by [27]. This simple, yet strong method will help researchers come up with new ideas, as well as simplify the evaluation.

The main contributions of this article are as follows:

- We propose a large-scale mouse pose dataset for mouse pose estimation. It makes up for the shortage of uniform and standardized datasets in mouse pose estimation.
- We design a fast and convenient keypoint annotation tool. The features of being easy to reproduce and employ make it have extensive potential applications in related work.
- A simple and efficient pipeline as a benchmark is proposed for evaluation on our dataset.

Our paper is organized as follows. In Section 2, we review the existing datasets of the mouse in the deep learning area and analyze their features. Related work in pose estimation is also presented in this section. In Section 3, we describe our capturing device used for collecting the data. Section 4 describes the dataset we propose in detail. Section 5 describes

our benchmark used for mouse pose estimation, including experimental networks, evaluation standards, experimental settings, and results. The paper ends with the conclusions, which is Section 6.

2. Related Work

2.1. Datasets for the Mouse Poses

There exists a range of 2D mouse pose datasets varying in multifarious aspects. Hu et al. [8] created a dataset of the mouse composed of 4627 frames of 2D poses (20 keypoints) from three sets of video data. The dataset was collected in the dark cycle with infrared illumination. Within this, 32 mice were distributed into four different classes. They were caged independently to capture mostly daily behaviors. The PDMB dataset [28] contains four videos of four mice, and each video was divided into six ten-minute clips with 9248 images used. Both of the datasets above were collected from real-world videos. Additionally, Xu et al. [29] provided 3253 depth images of two different lab mice, which successfully helped them acquire distinct poses, as well as depth noise patterns. However, unlike the above datasets, they tripled the size of the dataset through additional transformations. Another special dataset was released by Mu et al. [30]. This dataset is constituted by synthetic images based on the Coco val2017 dataset. Obviously, synthesis techniques are also becoming increasingly prevalent in the domain.

A set of systematic datasets of mice has also been proposed in recent years. The CalMS21 dataset was produced from raw 30 Hz videos [31]. It consists of not only six million frames of unlabeled tracked poses of interacting mice, but also over one million frames of tracked poses and corresponding frame-level behavior annotations (seven keypoints). Unfortunately, all the data of CalMS21 were designed to be targeted for studying behavior classifications. It does not match work in pose estimations to some extent. The Paired Acquisition of Interacting oRganisms–Rat (PAIR-R24M) dataset was prepared for multi-animal 3D pose estimations [32]. It contains 24.3 million frames of RGB video of 18 different pairs of laboratory rats (11 keypoints) from 24 viewpoints and 3 interaction categories. Dissecting a mass of existing mouse datasets, various problems are ubiquitous, including few research objects and unclear descriptions of the process of obtaining datasets [13,33].

Thus far, the need for 3D pose estimations is growing with the advancement of deep learning technologies. Two-dimensional images are not only key to their analyses, but also fundamental for further research [8,34–37]. However, after analyzing the datasets mentioned above, several universal limitations are obvious among existing mouse datasets: The collecting environments were not uniform, which largely limits the efficiency of employing the data; the datasets were collected for a specific target, but not for pose estimation, which does not match the pose estimation work; some datasets were created by transformation techniques, which are not real data.

Therefore, our work aims to provide a large-scale standardized and annotated mouse pose dataset. The data were collected from pairs of mice in a stable environment. Each image is very clear and high quality such that it can satisfy not only our work—the estimations of mice poses—but also, it can be easily utilized in other aspects of related research on the mouse based on deep learning technologies. Evidently, our dataset has a more extensive application prospect. More details about the dataset will be introduced in Section 4.

2.2. Annotating Software and Hardware Devices

Currently, the need for automated and efficient software for annotating pose images has sharply increased with massive images. At the beginning of the development of pose estimations, most image annotation was performed by humans [25,26]. This largely increases the cost and complexity of research. Under the necessity of relieving the work of humans, simple but effective annotating software has arisen in response to the time and conditions. Object detection with semi-annotated weak labels (ODSAWLs) needs the

image-level tags for a small portion of the training images [20]. It cooperates with object detectors, which can be learned from a small portion of the weakly labeled training images, as well as from the remaining unlabeled training images. Recently, DeepLabCut has been utilized in this field [38,39]. It is a method for markerless pose estimations based on transfer learning with minimal training data. On this basis, automatic software and devices are created to aid in freeing humans from these time-consuming tasks [40,41]. However, in order to amplify the application scope of our mouse dataset, we introduce a simple, but effective annotating software for fast, vigorous, and available image markers.

In parallel, it is common to find capturing devices set up in the field of pose estimations. They satisfy various requirements of the observation angles. Hsien et al. [25] built a hardware setup with a behavior apparatus, a sensor device, and a personal computer. Wang et al. [42] set up an experimental device for data acquisitions. Here, we also built a hardware device for data collection, illustrated in Section 3.

2.3. Algorithms and Baselines of Pose Estimation

Simple, yet effective baseline methods are beneficial to inspire and evaluate new ideas for the field [27]. Recent advances in human pose estimations have resulted in various baselines of human behavior being proposed. CPN [17] aimed to handle the keypoints that were occluded, invisible, or in a complex background through integrating all levels of the feature representations from the GlobalNet. Xiao ed al. [27] proposed a baseline that was validated to outperform other methods for 2D human pose estimation and tracking. Andriluka et al. [15] proposed two baselines that performed well on easy sequences with well-separated upright people; however, this is not well suited for fast camera motions and complex articulations. InterHand2.6M [43] contains both a dataset and a baseline for 3D interacting hand pose estimations from RGB images and built a solid foundation for future works. Marinez et al. [44] released a high-performance, yet lightweight and easy-to-reproduce baseline of 3D human pose estimations. Their work sets a bar for future works in this field. However, compared with the quick maturity of baselines in human pose estimations, simple and effective baselines of animal pose estimations need to be explored in the mouse pose estimation field.

3. Capturing Device

We designed a device suitable for the laboratory environment to collect the data. The device was used for data acquisitions of real-time pose information of the interacting mice, including a capturing apparatus, the Logitech C270 sensor camera, and a personal computer (Figure 1). To stabilize the equipment, a black steel plate was placed at the bottom of the alloy body. The capturing apparatus consisted of a cube metalbody (30 cm × 30 cm × 30 cm), two hinged rotating metalarms (140 cm), and a circular fill light modulator (r = 13 cm). The sensor camera was inserted into the center of our light modulator, mounted 80 cm above the steel plate at the bottom to obtain clear, accurate, and stable RGB image data of the mice. The two hinged rotating arms were fixed at approximately 130° and 165°, respectively, to provide consecutive stable video shooting. Both the height, as well as the angle can be adjusted at will.

The Logitech C270 camera provides high-quality images with a resolution of up to 720p. Despite it having the function of multi-person calls, we did not use this function, as we wanted to concentrate more on our precise data collections and minimize the negative performance effect while capturing the images. As shown in Figure 1, the camera was connected to a personal computer for recording the videos of the laboratory mice and storing them. The process of extracting the frames of RGB images from recorded videos was implemented on the computer, in which the sampling rate was set at 30 frames/s (30 Hz).

Furthermore, four Yagli boards were utilized to create a space for the movement of the mice. The boards not only guarantee the overall activities of the mice in the range of the customized capturing device, but also makes the environment closer to the biological

mouse laboratory. The experimental device was able to acquire abundant accurate video information on the activities, as well as the movements of the mice.

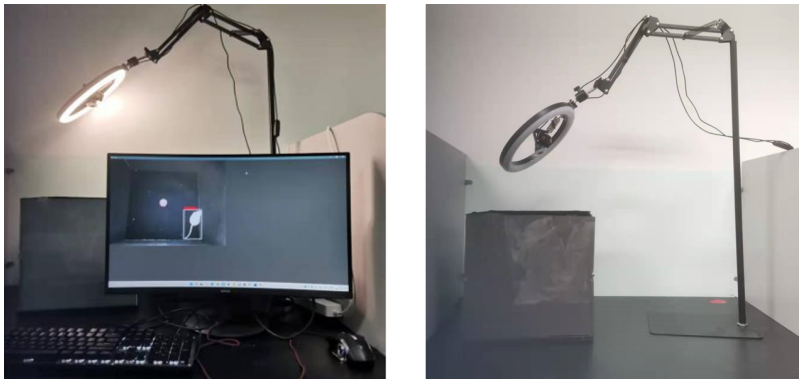

Figure 1. Capturing device.

4. Data Description

Our proposed mouse pose dataset was designed to provide abundant quantities of training data for mouse pose estimations. The dataset was structured into RGB images, mouse area locations, and 2D keypoint positions, for which each image was composed of captured frames under the rate of 30 frames/s (30 Hz). In particular, the composition of this dataset was as follows:

- A series of 2D RGB images of mice in the experimental setting.
- The bounding box for positioning the mouse in the image.
- Annotated mouse keypoint coordinates.

Additionally, the uniformity in the species, illumination, living environment, and observation angles profoundly ensured the reliability, as well as the quality of the mouse pose dataset. Controlling these variables will definitely make the mouse pose dataset have a well-directed and functional role in pertinent fields, advancing the efficiency of the primary work in machine learning.

4.1. Definitions of Mouse 2D Joint Points

For each frame in the dataset, a set of 2D points is provided. These two-dimensional points correspond to the keypoints of the mice in the laboratory environment, requiring no further preprocessing.

Table 1 lists the ID of each point and its semantic correspondence. In Figure 2, five different points are marked on one mouse with their corresponding X and Y coordinates. As the analysis above, we set up five keypoints based on experience [37]: mouth, left ear, right ear, neck, and tail root.

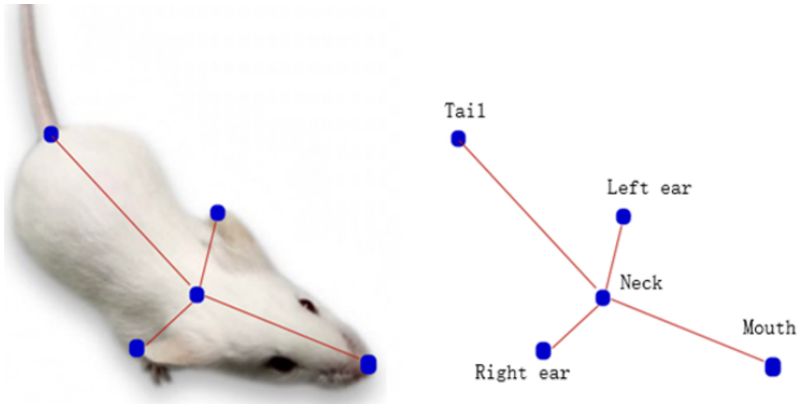

Figure 2. Five keypoints marked on one mouse with their corresponding X and Y coordinates.

Table 1. The ID of each keypoint of a mouse in the software and its semantic name.

Joint ID	Semantic Name
Tag 1	Mouth
Tag 2	Left Ear
Tag 3	Right Ear
Tag 4	Neck
Tag 5	Tail Root

4.2. Color Images of a Mouse

The dataset we created is mainly for mouse pose estimation systems based on deep learning, while other fields were also considered. Within all these systems, the acceptable loss of pose estimations is related to the quality of the input RGB mouse images. Therefore, the quality of the input images holds great importance at present. As stated before, every frame of the mouse pose dataset is a color image, which is recorded from a top-down view. Notably, there were slight deformations while the vision sensor was capturing the images. Fortunately, the camera we used has the ability to handle image distortions, which allowed the images to meet our requirements.

4.3. Mouse 2D Joint Point Annotations

In the past, the traditional method of capturing keypoints was to install sensors at the joints of humans or animals and obtain joint point coordinates by analyzing sensor data. However, it is very difficult to install sensors on the joint points of the body of small animals, especially mice. In this way, we chose to shoot active videos of these small animals at first. Then, we took the frames to obtain images and mark the joints of animals on the images. This method can overcome the problem of not being able to install sensors on small animals.

The keypoints of our dataset were the five most easily observable in the top-down perspective (Figure 3). At the same time, these five keypoints can simulate the daily behavior of most mice. Therefore, they can be well applied in the laboratory environment. To obtain the annotated 2D pose data of mice, we divided the annotation task into two parts. In the first part, we used the LabelImg application [45] to annotate the mouse locations. Then, we cut out the mouse images from the original images based on the mouse localization coordinates.

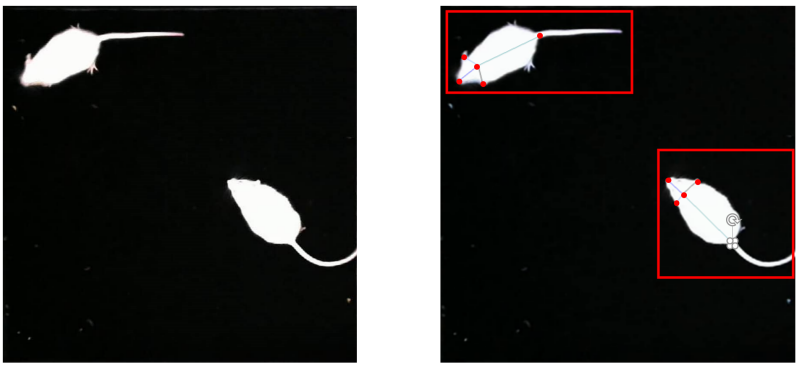

Figure 3. The top-down perspective of a mouse pose captured by the hardware device.

In the second part, we performed keypoint annotation on cropped mouse images. To facilitate the execution of keypoint annotations, we produced a universal mouse pose estimation labeling software (Figure 4). The software is based on PyQt5, a Python language implemented on the basis of the graphical programming framework Qt5, which consists of a set of Python modules. The PyQt5 API has more than 620 classes and 6000 functions. These well-packaged classes and functions make it easier and more convenient for users to instantiate classes and call functions. It is a cross-platform toolkit that can run on all

major operating systems, including Windows, Linux, and Mac OS. All the advantages shown above contributed to our choice of PyQt5 as the means to process the images. It can annotate not only the joints of mice, but also the joints of other animals in the image. At present, no labeling software on the market is specifically aimed at labeling the keypoints of objects in an image. Our self-created annotating software is based on the python3.6 and PyQt5 libraries. The basic functions of this software are to visualize the labeling process and save the coordinates of the annotated keypoints in a text document file. At the same time, in order to improve the efficiency of the labeling, we also added some functions that facilitate the labeling process, such as adding a quick interface, switching between multiple files, and removing labeling points.

Figure 4. The basic interface of the annotating software.

Finally, it is worth mentioning that the reason why we determined the top-down mouse pose capture perspective was to ensure that we could observe every joint point of the mouse without interfering with the daily activities of the mouse, which made our mouse pose estimation dataset more accurate.

4.4. Variability and Generalization Capabilities

Releasing our dataset of mouse pose estimations is for the purpose of providing high-precision ground-truth data. However, the progress was hindered by the characteristics of mouse activities, which are autonomous, uncontrolled, and unscheduled. This is mainly due to individual differences: a large proportion of experimental mice with independent, yet unfixed postures will be obscured by their bodies. In parallel, exceptional cases also occurred in the course of continuous observations. For example, multiple mice overlapped each other. Therefore, in the process of labeling, eight skilled annotators were engaged, and they manually checked, as well as eliminated such unqualified data. Specifically, when the feature points in the image were covered by other parts of the body, we directly deleted such data to ensure the correctness and validity of the dataset. Furthermore, cross-checking was applied to the examination process of the annotated dataset, effectively avoiding artificial errors. Every mouse in our laboratory was a healthy and normal individual.

To this end, we used multiple mice for video data acquisitions in different permutations and combinations and excluded those frames that were clustered together. In conclusion, our mouse pose dataset contains 40,000 2D RGB images of mice living in the laboratory environment. Profiting from the manual elaboration, each image of the dataset can thoroughly represent the pose of a mouse. With the need to generate training data and test segmentation data, the mouse pose dataset was recombined, and 20% served for testing, while the remaining 80% were for training.

5. Benchmark—2D Keypoint Estimations

In this section, we propose a benchmark model based on deep learning algorithms, which includes the process of mouse detection, mouse pose estimation, the evaluation standard, the experimental settings, and the experimental results. To this end, a pipeline from mouse images to 2D keypoints is proposed.

5.1. Mouse Detection

First, our detection device utilizes a Logitech C270 camera to record video segments of mice and arranges the video into a series of RGB images at a constant 30 frames per second rate. In the second part, all eligible data are transported through the trained network YOLOv4 [46], which is applied to determine the locations of mice that appeared in the scene. The YOLOv4 network structure is shown in Figure 5.

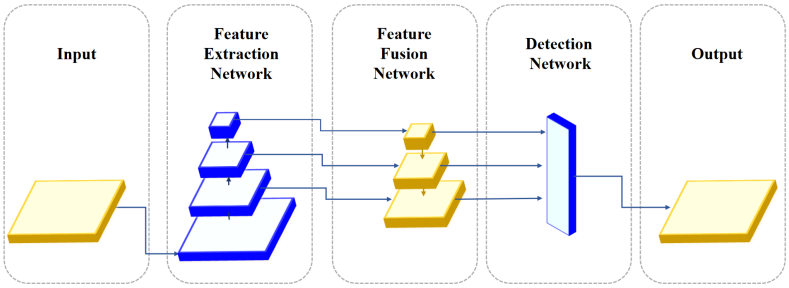

Figure 5. The structure of the YOLOv4 network.

YOLOv4 has a relatively big change compared to YOLOv3. First, the original Leaky-ReLU is replaced by the Mish function in the network structure of feature extraction, as shown in Equation (1).

$$Mish = x \times tanh(ln(1 + e^x)) \quad (1)$$

This change guarantees the flow of information while ensuring that negative values are not completely truncated, thereby avoiding the problem of gradient saturation. At the same time, the *Mish* function compared with ReLU also makes sure there is no smoothing effect, making the effect of gradient descent better than that of ReLU. In the equation, x represents the pixels of the input image; the outputs of YOLOv4 include both the bounding box of the mice and the score representing the detection confidence.

5.2. Mouse Pose Estimation

Mouse pose estimation is the third process of our benchmark. Within this process, each image of the mice is cropped based on the output of YOLOv4 and is adjusted to 256 × 256 pixels. It is fed to the 2D pose estimation network [27] for themouse keypoint coordinates. We found that the best choice was Adam, whose learning rate was 0.003. The loss function we used was the MSE. This is an end-to-end process. The overall pipeline is displayed in Figure 6.

Our baseline method was verified in the test, which was processed with test segmentation cross-validation, and the average absolute error of validation for 256 × 256 mouse images was 0.02%, i.e., 10-pixel error. The results based on the real image data were also acquired in the experiment, which will be presented in Section 5.5.

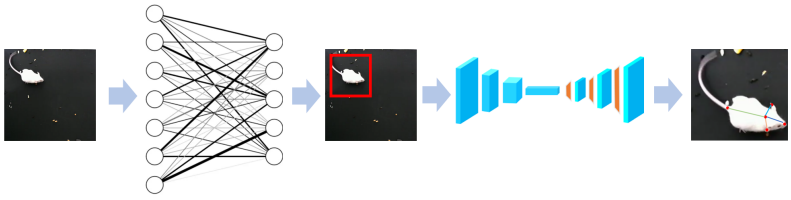

Figure 6. The structure of the pipeline.

Moreover, due to the single video background and controllable external disturbances, the operation of pruning the network of pipelines properly was very beneficial. For example, we used a backbone network with fewer parameters. That not only reduced the cost of the computation during training, but also promoted the efficiency of mouse pose estimation.

5.3. Evaluation Standard

Our baseline model consists of two parts, object detection and pose estimation. In the object detection part, the images in the test set are input into the algorithm. If the intersection over union (IOU) of the bounding box of the mouse detected in the test image and the bounding box in the label is greater than or equal to the threshold, we set (0.6); the mice were considered to be successfully detected. In this paper, the accuracy rate (precision (P)) was used as the evaluation index of the accuracy of the target detection model. The calculation formula is as follows:

$$P = \frac{TP}{TP + FP} \tag{2}$$

In Equation (2), TP indicates the number of correctly detected mice in the test set; FP indicates the number of falsely detected mice in the test set. In the pose estimation part, the percentage of correct keypoints (PCK) was used as the average error in each keypoint and label data to evaluate the effect of the algorithm in pixels.

5.4. Experimental Settings

In this section, we gradually introduce our experimental environment and pose estimation results from the configuration of the experiment.

All the results of our pose estimations were obtained by experiments with the following experimental equipment: Ubuntu 20.04 as the operating system of the experiment, Pytorch 1.6 as the deep learning framework used in all experiments, and an NVIDIA Geforce RTX 2080s GPU, with a video memory of 8 GB, from which all experimental results were obtained.

In the pose estimation process, the total pose was estimated to run at 27 frames per second and can be tuned in the code to run at 30 frames per second or 15 frames per second. In the object detection process, we used 30 frames per second. For example, on the NVIDIA Geforce RTX 2080, the mouse pose was estimated to take only 10 ms per frame. Our model framework was initially trained and tested on the COCO dataset [47], running on Ubuntu20.04, using CMake 3.16.3, GCC 7.5.0, CUDA 11.4, and cuDNN 8.24.

5.5. Experimental Results

In the mouse detection experiment, it is worth noting that we trained the YOLOv4 network independently. For the purpose of improving the efficiency and relevance of the experiment, we actively selected the output parameters, which were all required by the experiment not only when evaluating experiments, but also when demonstrating baseline performance. Thus, there no suspicious parameters needed to be excluded. During the process, there were 7844 ground-truth images, among which 7535 images were successfully detected. They were the input of the Yolov4 network. With the rate of 30 frames per second in the training procedure, the counting accuracy was 0.96 and the average precision was 0.91. Table 2 shows the relevant parameters of our object detection experiment for training the YOLOv4 network.

Table 2. The relevant data on the experiment of object detection.

Item	Object Detection
Ground Truth	7844
Detected	7535
Average Precision	0.91
Counting Accuracy	0.96
Frames Per Second	30

When it comes to the mouse pose estimation experiment, there were 37,502 ground-truth real images used as the input of the pose estimation network. Since our experimental parameters were not complicated and our method was to actively choose the parameters, all the output parameters were essential. With the rate of 27 frames per second in this

procedure, the percentage of correct keypoints was 85%. Table 3 shows the relevant parameters of our pose estimation experiment.

Table 3. The relevant data on the experiment of mouse pose estimation.

Item	Pose Estimation
Ground-Truth	37,502
Percentage of Correct Keypoints (PCK)	85%
Frames Per Second	27

The evaluation results of our experiments are shown in Table 4. The high accuracy of the mouse object detection was due to the fact that our object was specific, that is mice, with less background noise, so even if we used a small-scale network, we could achieve a high-accuracy detection. The percentage of correct keypoints in pose estimation was 85%, which still needs to be improved in future experiments.

Table 4. The evaluation results of the object detection and pose estimation experiments.

Method	Intersection over Union (IOU)	Percentage of Correct Keypoints (PCK)
Object Detection	0.9	\
Pose Estimation	\	85%

6. Conclusions

We introduced a mouse pose dataset, a novel dataset with each image annotated to estimate the keypoints of mice in a laboratory setting. The proposed mouse pose dataset is the first standardized large-scale 2D mouse pose dataset and involves 40,000 single and interacting mouse images from pairs of laboratory mice. A creative software for annotating the images was produced, which largely frees humans from the time-consuming work. In addition, a simple, yet effective baseline was provided here using the deep learning network. Our dataset provides a solid guarantee for various potential future applications on animal pose estimations. In future work, we will continue to expand our dataset from 2D mouse poses to 3D mouse poses. At the same time, we will try to introduce newer methods, such as self-supervised and unsupervised methods, to achieve better 2D and 3D pose estimations of mice.

Author Contributions: Conceptualization, J.S.; methodology, J.S.; software, X.L. and S.W.; validation, J.S. and M.W.; formal analysis, J.S. and M.W.; investigation, J.S. and J.W.; resources, M.W.; writing—original draft preparation, J.S., J.W. and X.L.; writing—review and editing, J.S. and M.W.; visualization, X.L. and S.W.; supervision, M.W.; project administration, J.S. and M.W.; funding acquisition, M.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Sichuan Agricultural University (Grant No. 202110626117, 202010626008).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors thank the anonymous Reviewers for the helpful comments, which improved this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Sample Availability: The dataset link is: https://github.com/lockeding/Mouse-Resource (accessed on 1 March 2022).

References

1. Lewejohann, L.; Hoppmann, A.M.; Kegel, P.; Kritzler, M.; Krüger, A.; Sachser, N. Behavioral phenotyping of a murine model of alzheimer's disease in a seminaturalistic environment using rfid tracking. *Behav. Res. Methods* **2009**, *41*, 850–856. [CrossRef] [PubMed]
2. Geuther, B.Q.; Peer, A.; He, H.; Sabnis, G.; Philip, V.M.; Kumar, V. Action detection using a neural network elucidates the genetics of mouse grooming behavior. *Elife* **2021**, *10*, e63207. [CrossRef] [PubMed]
3. Hutchinson, L.; Steiert, B.; Soubret, A.; Wagg, J.; Phipps, A.; Peck, R.; Charoin, J.E.; Ribba, B. Models and machines: How deep learning will take clinical pharmacology to the next level. *CPT Pharmacomet. Syst. Pharmacol.* **2019**, *8*, 131. [CrossRef]
4. Ritter, S.; Barrett, D.G.; Santoro, A.; Botvinick, M.M. Cognitive psychology for deep neural networks: A shape bias case study. In Proceedings of the International Conference on Machine Learning (PMLR 2017), Sydney, Australia, 6–11 August 2017; pp. 2940–2949.
5. Fang, H.-S.; Xie, S.; Tai, Y.-W.; Lu, C. Rmpe: Regional multi-person pose estimation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2334–2343.
6. Supancic, J.S.; Rogez, G.; Yang, Y.; Shotton, J.; Ramanan, D. Depth-based hand pose estimation: Data, methods, and challenges. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1868–1876.
7. Toshev, A.; Szegedy, C. Deeppose: Human pose estimation via deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1653–1660.
8. Hu, B.; Seybold, B.; Yang, S.; Ross, D.; Sud, A.; Ruby, G.; Liu, Y. Optical mouse: 3d mouse pose from single-view video. *arXiv* **2021**, arXiv:2106.09251.
9. Li, X.; Cai, C.; Zhang, R.; Ju, L.; He, J. Deep cascaded convolutional models for cattle pose estimation. *Comput. Electron. Agric.* **2019**, *164*, 104885. [CrossRef]
10. Badger, M.; Wang, Y.; Modh, A.; Perkes, A.; Kolotouros, N.; Pfrommer, B.G.; Schmidt, M.F.; Daniilidis, K. 3d bird reconstruction: a dataset, model, and shape recovery from a single view. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 1–17.
11. Psota, E.T.; Mittek, M.; Pérez, L.C.; Schmidt, T.; Mote, B. Multi-pig part detection and association with a fully-convolutional network. *Sensors* **2019**, *19*, 852. [CrossRef]
12. Sanakoyeu, A.; Khalidov, V.; McCarthy, M.S.; Vedaldi, A.; Neverova, N. Transferring dense pose to proximal animal classes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5233–5242.
13. Pereira, T.D.; Aldarondo, D.E.; Willmore, L.; Kislin, M.; Wang, S.S.; Murthy, M.; Shaevitz, J.W. Fast animal pose estimation using deep neural networks. *Nat. Methods* **2019**, *16*, 117–125. [CrossRef] [PubMed]
14. Behringer, R.; Gertsenstein, M.; Nagy, K.V.; Nagy, A. *Manipulating the Mouse Embryo: A Laboratory Manual*, 4th ed.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, USA, 2014.
15. Andriluka, M.; Iqbal, U.; Insafutdinov, E.; Pishchulin, L.; Milan, A.; Gall, J.; Schiele, B. Posetrack: A benchmark for human pose estimation and tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5167–5176.
16. Andriluka, M.; Pishchulin, L.; Gehler, P.; Schiele, B. 2d human pose estimation: New benchmark and state of the art analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3686–3693.
17. Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; Sun, J. Cascaded pyramid network for multi-person pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7103–7112.
18. Insafutdinov, E.; Pishchulin, L.; Andres, B.; Andriluka, M.; Schiele, B. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 34–50.
19. Iqbal, U.; Milan, A.; Gall, J. Posetrack: Joint multi-person pose estimation and tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2011–2020.
20. Tompson, J.J.; Jain, A.; LeCun, Y.; Bregler, C. Joint training of a convolutional network and a graphical model for human pose estimation. *Adv. Neural Inf. Process. Syst.* **2014**, *27*. [CrossRef]
21. Liu, X.; Yu, S.-Y.; Flierman, N.; Loyola, S.; Kamermans, M.; Hoogland, T.M.; De Zeeuw, C.I. Optiflex: Video-based animal pose estimation using deep learning enhanced by optical flow. *BioRxiv* **2020**. [CrossRef]
22. Machado, A.S.; Darmohray, D.M.; Fayad, J.; Marques, H.G.; Carey, M.R. A quantitative framework for whole-body coordination reveals specific deficits in freely walking ataxic mice. *Elife* **2015**, *4*, e07892. [CrossRef] [PubMed]
23. Marks, M.; Qiuhan, J.; Sturman, O.; von Ziegler, L.; Kollmorgen, S.; von der Behrens, W.; Mante, V.; Bohacek, J.; Yanik, M.F. Deep-learning based identification, pose estimation and end-to-end behavior classification for interacting primates and mice in complex environments. *bioRxiv* **2021**. [CrossRef]
24. Pereira, T.D.; Tabris, N.; Li, J.; Ravindranath, S.; Papadoyannis, E.S.; Wang, Z.Y.; Turner, D.M.; McKenzie-Smith, G.; Kocher, S.D.; Falkner, A.L.; et al. Sleap: Multi-animal pose tracking. *BioRxiv* **2020**. [CrossRef]

25. Ou-Yang, T.H.; Tsai, M.L.; Yen, C.-T.; Lin, T.-T. An infrared range camera-based approach for three-dimensional locomotion tracking and pose reconstruction in a rodent. *J. Neurosci. Methods* **2011**, *201*, 116–123. [CrossRef] [PubMed]
26. Hong, W.; Kennedy, A.; Burgos-Artizzu, X.P.; Zelikowsky, M.; Navonne, S.G.; Perona, P.; Anderson, D.J. Automated measurement of mouse social behaviors using depth sensing, video tracking, and machine learning. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, E5351–E5360. [CrossRef] [PubMed]
27. Xiao, B.; Wu, H.; Wei, Y. Simple baselines for human pose estimation and tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 466–481.
28. Zhou, F.; Jiang, Z.; Liu, Z.; Chen, F.; Chen, L.; Tong, L.; Yang, Z.; Wang, H.; Fei, M.; Li, L.; et al. Structured context enhancement network for mouse pose estimation. *IEEE Trans. Circuits Syst. Video Technol.* **2021**. [CrossRef]
29. Xu, C.; Govindarajan, L.N.; Zhang, Y.; Cheng, L. Lie-x: Depth image based articulated object pose estimation, tracking, and action recognition on lie groups. *Int. J. Comput. Vis.* **2017**, *123*, 454–478. [CrossRef]
30. Mu, J.; Qiu, W.; Hager, G.D.; Yuille, A.L. Learning from synthetic animals. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12386–12395.
31. Sun, J.J.; Karigo, T.; Chakraborty, D.; Mohanty, S.P.; Wild, B.; Sun, Q.; Chen, C.; Anderson, D.J.; Perona, P.; Yue, Y.; et al. The multi-agent behavior dataset: Mouse dyadic social interactions. *arXiv* **2021**, arXiv:2104.02710.
32. Marshall, J.D.; Klibaite, U.; Gellis, A.J.; Aldarondo, D.E.; Olveczky, B.P.; Dunn, T.W. The pair-r24m dataset for multi-animal 3d pose estimation. *bioRxiv* **2021**. [CrossRef]
33. 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. Multi-animal pose estimation and tracking with deeplabcut. *BioRxiv* **2021**. [CrossRef]
34. Günel, S.; Rhodin, H.; Morales, D.; Campagnolo, J.; Ramdya, P.; Fua, P. Deepfly3d, a deep learning-based approach for 3d limb and appendage tracking in tethered, adult drosophila. *Elife* **2019**, *8*, e48571. [CrossRef]
35. Mathis, M.W.; Mathis, A. Deep learning tools for the measurement of animal behavior in neuroscience. *Curr. Opin. Neurobiol.* **2020**, *60*, 1–11. [CrossRef] [PubMed]
36. Salem, G.; Krynitsky, J.; Hayes, M.; Pohida, T.; Burgos-Artizzu, X. Three-dimensional pose estimation for laboratory mouse from monocular images. *IEEE Trans. Image Process.* **2019**, *28*, 4273–4287. [CrossRef]
37. Nanjappa, A.; Cheng, L.; Gao, W.; Xu, C.; Claridge-Chang, A.; Bichler, Z. Mouse pose estimation from depth images. *arXiv* **2015**, arXiv:1511.07611.
38. Mathis, A.; Mamidanna, P.; Cury, K.M.; Abe, T.; Murthy, V.N.; Mathis, M.W.; Bethge, M. Deeplabcut: Markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* **2018**, *21*, 1281–1289. [CrossRef] [PubMed]
39. Nath, T.; Mathis, A.; Chen, A.C.; Patel, A.; Bethge, M.; Mathis, M.W. Using deeplabcut for 3d markerless pose estimation across species and behaviors. *Nat. Protoc.* **2019**, *14*, 2152–2176. [CrossRef]
40. Graving, J.M.; Chae, D.; Naik, H.; Li, L.; Koger, B.; Costelloe, B.R.; Couzin, I.D. Deepposekit, a software toolkit for fast and robust animal pose estimation using deep learning. *Elife* **2019**, *8*, e47994. [CrossRef] [PubMed]
41. Zhang, Y.; Park, H.S. Multiview supervision by registration. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Seattle, WA, USA, 14–19 June 2020; pp. 420–428.
42. Wang, Z.; Mirbozorgi, S.A.; Ghovanloo, M. An automated behavior analysis system for freely moving rodents using depth image. *Med. Biol. Eng. Comput.* **2018**, *56*, 1807–1821. [CrossRef] [PubMed]
43. Moon, G.; Yu, S.; Wen, H.; Shiratori, T.; Lee, K.M. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 548–564.
44. Martinez, J.; Hossain, R.; Romero, J.; Little, J.J. A simple yet effective baseline for 3d human pose estimation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2640–2649.
45. TzuTa Lin. Labelimg. 2015. Available online: https://github.com/tzutalin/labelImg (accessed on 1 March 2022).
46. Bochkovskiy, A.; Wang, C.; Liao, H.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
47. Lin, T.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.