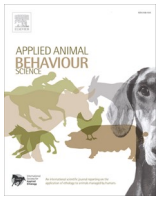
Applied Animal Behaviour Science 277 (2024) 106347

Contents lists available at ScienceDirect

Applied Animal Behaviour Science

journal homepage: www.elsevier.com/locate/applanim



Skeleton-based image feature extraction for automated behavioral analysis in human-animal relationship tests

Maciej Oczak [a,b,*], Jean-Loup Rault [b], Suzanne Truong [b], Oceane Schmitt [b,c]

[a] Precision Livestock Farming Hub, The University of Veterinary Medicine Vienna (Vetmeduni Vienna), Veterinärplatz 1, Vienna 1210, Austria
[b] Center for Animal Nutrition and Welfare, The University of Veterinary Medicine Vienna (Vetmeduni Vienna), Veterinärplatz 1, Vienna 1210, Austria
[c] Chair Animal Husbandry, Behaviour and Welfare, Institute of Animal Breeding and Genetics, Justus-Liebig University of Giessen, Leihgesterner Weg 52, Giessen 35392, Germany

ARTICLE INFO

Keywords:
Object detection
Key point detection
Human-animal
Distance
Computer vision

ABSTRACT

Arena tests are used to address various research questions related to animal behavior and human-animal relationships; e.g. how animals perceive specific human beings or people in general. Recent advancements in computer vision, specifically in application of key point detection models, might offer a possibility to extract variables that are the most often recorded in these tests in an automated way. The objective of this study was to measure two variables in human-pig arena test with computer vision techniques, i.e. distance between the subjects and pig's visual attention proxy towards pen areas including a human. Human-pig interaction tests were organized inside a test arena measuring 147 × 168 cm. Thirty female pigs took part in the arena tests from 8 to 11 weeks of age, for a total of 210 tests (7 tests per pig), each with a 10-min duration. In total, 35 hours of human-pig interaction tests were video-recorded. To automatically detect human and pig skeletons, 4 models were trained on 100 images of labeled data, i.e. two YOLOv8 models to detect human and pig locations and two VitPose models to detect their skeletons. Models were validated on 50 images. The best performing models were selected to extract human and pig skeletons on recorded videos. Human-pig distance was calculated as the shortest Euclidean distance between all key points of the human and the pig. Visual attention proxy towards selected areas of the arena were calculated by extracting the pig's head direction and calculating the intersection of a line indicating the heads direction and lines specifying the areas i.e. either lines of the quadrangles for the entrance and the window or lines joining the key points of the human skeleton. The performance of the YOLOv8 for detection of the human and the pig was 0.86 mAP and 0.85 mAP, respectively, and for the VitPose model 0.65 mAP and 0.78 mAP, respectively. The average distance between the human and the pig was 31.03 cm (SD = 35.99). Out of the three predefined areas in the arena, pigs spend most of their time with their head directed toward the human, i.e. 12 hrs 11 min (34.83 % of test duration). The developed method could be applied in human-animal relationship tests to automatically measure the distance between a human and a pig or another animal, visual attention proxy or other variables of interest.

1. Introduction

In controlled experiments, animals are typically placed in a square or round arena, open on the top and limited by walls on the sides, in order to assess their reactivity to a novel environment or stimulus (Grabovskaya and Salyha, 2014). Arena tests are used to address various research questions related to animal behavior and human-animal relationships; e.g. how animals perceive specific human beings or people in general, what is the quality of stockmanship on farms or what is the potential for positive relationships to reduce the animals' distress during aversive events (Waiblinger et al., 2006). The variables that are the most often recorded in these tests and analyzed to assess the behavior are: distance to the human, latency to contact, frequency or duration of contact, visual attention towards the human or locomotor activity such as lying, standing, or exploration (Bensoussan et al., 2020; Czycholl et al., 2019; Forkman et al., 2007). Recent advancements in computer vision, specifically in application of key point detection models for automated detection of pig's skeletons e.g Wang et al. (2022) or Wang

* Corresponding author at: Precision Livestock Farming Hub, The University of Veterinary Medicine Vienna (Vetmeduni Vienna), Veterinärplatz 1, Vienna 1210, Austria.
E-mail address: Maciej.Oczak@vetmeduni.ac.at (M. Oczak).

https://doi.org/10.1016/j.applanim.2024.106347
Received 6 May 2024; Received in revised form 25 June 2024; Accepted 1 July 2024
Available online 2 July 2024

et al. (2023) might offer a possibility to extract those variables in an automated way, reducing the need for laborious manual labeling on a farm or of recorded videos.

Skeleton-based deep learning models were first developed to detect key points in humans, which required overcoming challenges such as strong articulations, small and barely visible joints, occlusions and the need to capture the context (Toshev and Szegedy, 2013). Recent generation of key point detection models, e.g. the ViTPose model, achieves impressive results on the challenging MS COCO Keypoint Detection benchmark setting a new state-of-the-art, i.e. 80.9 mean average precision (mAP) on the MS COCO test-dev set on a human skeleton with 17 key points (Xu et al., 2022). In the study of Juarez et al. (2023), the same model achieved 0.82 mAP on a comprehensive pig farm dataset with 5016 pig instances representing 6 production groups and encompassing the recognition of 22 key points on the pig skeleton. These results proved that both human and pig skeletons can be accurately detected with the current state-of-the-art key point detection models. Moreover, automated detection of location of key points or body parts was previously used to measure distances to the other objects in a pen or of pig's body parts e.g. as in studies of Ling et al. (2022), or Oczak et al. (2022), which suggests that automated analysis of animal-based variables in arena tests is a promising area of study.

In this study we propose a novel computer vision method for automated detection of variables in human-animal relationship tests, using the pig as a model. The method is based on a two-step approach where in the first step we detect the pig's and human's key points, and then based on the detected key points we extract two variables: the pig-human distance and the pig's visual attention proxy towards pen areas and a human inside the test arena. The applications of our method offer a possibility to extract those variables in an automated way, reducing the need for time-consuming manual observation or labeling of videos. It can also support objective analysis of animal behavior and human-animal relationships.

2. Material and methods

2.2. Animals, housing and human-animal interaction test

This study was performed as part of a larger experiment (manuscript in preparation) aiming to assess the effects of dopamine and opioid receptor antagonists on the behavior of the pigs interacting with a human. Two of the measures of interest in this overarching experiment were the distance between the pig and the human across the different test conditions, and the visual attention proxy of the pig towards the human, which were obtained through the elaboration of the algorithm presented in this paper.

Thirty female pigs (Sus scrofa domesticus; Swiss Large White × Pietrain breed) were used in this project, across two batches (batch 1 n= 20 pigs, batch 2 n= 10 pigs). The pigs were selected at weaning at the age of 5 weeks, based on health condition and avoiding weight extremities. The pigs were 5 weeks old at the start of habituation to the test pen and human, 8 weeks old at the start of testing in the arena, and 11 weeks old at the end of the experiment. Pigs were tested every third day. The remaining 2-days were "rest" days when pigs only received routine care.

The pigs were housed in two groups of 10 (batch 1) or 9 (batch 2). All pigs were tested, but only 5 pigs from each group were tested in the second batch. Pigs were housed in weaner pens measuring 2.45 × 3.82 m. Enrichment was provided in the form of two braided jute ropes of 1 m long hung from the pen fixtures in the lying area and two orange dog toy balls (Airflow ball, Dog Crest, 7.6 cm diameter).

The test pen consisted of two enclosures of similar size measuring 147 × 168 cm, made of yellow wood walls. One of the enclosures was the test arena, i.e. where the test pig was alone with the human, and the second one was a companion arena containing two companion pigs to minimize the effect of social separation. The partition between the two enclosures had a social window with dimensions of 50 × 25 cm allowing visual, auditory and olfactory contact between the test pig and the companion pigs (Fig. 1). Non-toxic animal paint was used to mark the pig on its back to allow individual recognition.

The interacting human remained seated in the corner opposite to the pen door for the whole 10-min duration of the test. The pig was encouraged to approach through vocal and physical solicitation cues, i.e. soft talking voice and tapping of the fingers, small hand and arm gestures but avoiding large and fast movements. The pig was allowed to voluntarily approach and interact with the human. If the pig was within arm's reach, the human provided gentle tactile contact, i.e. stroking, rubbing, scratching.

2.3. Video recording

The behavior of the pig and the human was video-recorded with one two-dimensional (2D) video camera (DS 2CD5046G0-AP; HikVision, Hangzhou, China) locked in protective housing HEB32K1 (Videotec, Schio, Italy) hanging above the arena in top view, 2 m above the pen. The images were recorded with a 1280 × 720 pixel resolution in MPEG-4 format, at 25 fps. The camera was connected to a server for the storage of video data (Synology, Taipei, Taiwan) with 4 cores, 8 GB memory, and 260 TB storage. The behavior of the human and the pig was video recorded for 10 min from the time when a pig entered the arena. In total, there were 210 tests divided into 2 batches with 140 tests in the first batch and 70 tests in the second batch. In total, 35 h of video recordings (210 videos x 10 min) were recorded during the experiments.

2.4. Camera calibration

To measure the distance between a pig and a human inside the test arena, in the units of the metric system of measurement e.g. centimeters, it was necessary to calibrate the camera used to record the tests. Initially, the recorded videos were cropped to 450 × 530 pixel resolution to remove the area outside of the test arena (Fig. 2a). In the second step 16 reference points were placed on the floor of the test arena, with known distances to each other (Fig. 2b). The points were placed on the floor when the arena was empty and removed before the tests started. These points' locations in the image (x,y) were subsequently used to obtain the camera intrinsic matrix and distortion coefficients according to Eqs. 1 and 2,

$$\textit{camera matrix} = \begin{bmatrix} 313.51 & 0 & 265.47 \\ 0 & 298.97 & 274.1 \\ 0 & 0 & 1 \end{bmatrix} \quad (1)$$

$$\textit{Distortion coefficients} = (0.03, -0.65, 0.01, -0.08, 0.51) \quad (2)$$

with the calibrateCamera function implemented in OpenCV (Bradski,

Fig. 1. Test arena. Test pig was placed in the test arena on the right with the interacting human during the test session. Two companion pigs were placed in the companion pen arena on the left during the test session and were provided access to food, water and straw.

Fig. 2. Camera calibration a) image cropped b) markers with known locations placed inside the arena c) image with corrected distortions - an empty arena d) image with corrected distortions - a pig interacting with a human.

2000).

With the known parameters of the camera intrinsic matrix and distortion coefficients the radial and tangential distortions in the images were corrected with the undistort function implemented in OpenCV (Figs. 2c and 2d).

Based on the camera calibration procedure, the size of one pixel in the video was estimated to 5.9 mm. All 210 videos, each of 10 min duration, recorded during the tests were undistorted according to this procedure.

2.5. Dataset for training of key point detection models

Out of 3150,000 frames recorded in the experiments (210 videos x 10 min x 60 s x 25 fps), 150 were selected with the K-means algorithm to guarantee that the selected images have the least correlation, as described in Pereira et al. (2019) and Oczak et al. (2023). Out of 150 images 100 were randomly selected as the training set and 50 as validation set for later training and validation of object detection and pose estimation models (Table 1).

Table 1
Dataset.

Dataset	N. images	N. humans	N. pigs
Training	100	73	67
Validation	50	37	28
Total	150	110	95

Out of 150 images 40 had no human and 55 no pig. Leaving images without either a pig or a human or both was intentional to provide negative examples for later training of object detection models i.e. without the presence of objects of interest.

2.6. Data labelling

The dataset with 150 images was labeled using the COCO annotator software V0.11.1, (Brooks, 2019). We modified the pig skeleton proposed by Juarez et al. (2023) described in Table 2 by removing the following 6 key points and their corresponding connections: tips of the

Table 2
Definitions of key points (Juarez et al. 2023).

Key point	Definition	Connected to
Snout	At the rostral bone in the tip of the nose, which comprises the parts of the face rostral to the eyes and dorsal to the mouth.	Neck
Base of the left ear	At the point where the left pinna, which is the portion of the ear that is visible on the outside of the head, connects to the skull.	Neck Tip of left ear
Base of the right ear	At the point where the right pinna, which is the portion of the ear that is visible on the outside of the head, connects to the skull.	Neck Tip of right ear
Tip of the left ear	At the distal end of the left pinna.	Base of left ear
Tip of the right ear	At the distal end of the right pinna.	Base of right ear
Neck	At the last cervical vertebra, which immediately precedes the thoracic vertebrae and marks the transition from the neck to the trunk.	Snout Base of left ear Base of right ear Right shoulder Left shoulder
Left shoulder	At the left scapulohumeral joint, the point where the articular head of the humerus connects with the scapula.	Neck Left flank Left elbow
Right shoulder	At the right scapulohumeral joint, the point where the articular head of the humerus connects with the scapula.	Neck Right flank Right elbow
Left elbow	At the left olecranon process, also known as the point of the elbow. It is found proximal and caudal to the elbow joint, where the distal end of the humerus articulates with the proximal ends of the radius and ulna.	Left shoulder Left hand
Right elbow	At the right olecranon process, also known as the point of the elbow. It is found proximal and caudal to the elbow joint, where the distal end of the humerus articulates with the proximal ends of the radius and ulna.	Right shoulder Right hand
Left hand	At the toe or tip of the hoof of the left front limb.	Left elbow
Right hand	At the toe or tip of the hoof of the right front limb.	Right elbow
Left flank	Between the left shoulder and the iliac bone of the left hip, in the centre of the left external abdominal oblique muscle, at the broadest part of the abdomen.	Left shoulder Left hip
Right flank	Between the right shoulder and the iliac bone of the right hip, in the centre of the right external abdominal oblique muscle, at the broadest part of the abdomen.	Right shoulder Right hip
Left hip	At the left coxofemoral joint, where the left femoral head meets the acetabulum of the os coxae.	Left flank Left knee Base of the tail
Right hip	At the right coxofemoral joint, where the right femoral head meets the acetabulum of the os coxae.	Right flank Right knee Base of the tail
Left knee	At the stifle joint, where the distal end of the left femur articulates with the tibia and the patella.	Left hip Left foot
Right knee	At the stifle joint, where the distal end of the right femur articulates with the tibia and the patella.	Right hip Right foot
Left foot	At the toe or tip of the hoof of the left hind limb.	Left knee
Right foot	At the toe or tip of the hoof of the right hind limb.	Right knee
Base of the tail	At the point where the first coccygeal vertebra meets the sacrum.	Left hip Right hip Tip of the tail
Tip of the tail	At the last coccygeal vertebra.	Base of the tail

ears, elbows and knees.

The skeleton was reduced to 16 key points with 17 connections (Fig. 3). This was done to reduce the complexity of the skeleton proposed by Juarez et al. (2023) and only use the key points that were considered the most essential for our study i.e. for calculation of pig-human distance and pig's head direction. To achieve these objectives it was not necessary to detect elbows, knees and 4 key points on the ears i.e. tips of the ears were removed, while the base of the ears were kept.

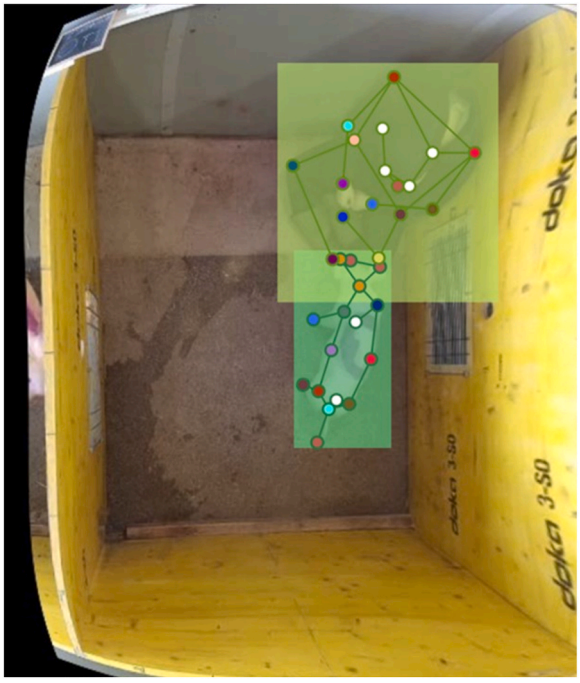


Fig. 3. Pig and human skeleton labeled in COCO annotator software with rectangles indicating a human and a pig.

Labeling of human skeleton was done with the COCO format i.e. with 17 key points and 17 connections (Lin et al., 2014).

2.7. Training and validation of the VitPose and the YOLOv8 models

To detect key body points of the human and the pig inside the test arena we applied a top down key point detection method. The method consisted of two steps: (1) detection of the pig and human with YOLOv8 as an object detection model, and (2) detection of key points with the VitPose key point detection model. Key points were detected only inside areas indicated as containing either a human or a pig by the object detection model. Both the VitPopse and the YOLOv8 models are state-of-the-art key point and object detection models that set new benchmarks on the MS COCO dataset (Solawetz, 2023; Xu et al., 2022). Moreover, the VitPose model was already validated for key point detection in pigs (Oczak et al., 2023). For training and validation of both models we used their implementations in the OpenMMLab toolbox (mmcv 2.0.1; mmdet 3.1.0; mmengine 0.8.4, mmyolo 0.6.0), which is an open source project containing implementations of state-of-the-art computer vision algorithms for object detection, animal pose estimation, action recognition, and tracking (MMDetection Contributors., 2018). Parametrization of the YOLOv8-x and the ViTPose-H model was set according to the implementation in the MMYolo and the MMPose libraries.

We trained 4 models: (1) the YOLOv8 for detection of the human, (2) the YOLOv8 for detection of the pig, (3) the VitPose for detection of human key points, and (4) the VitPose for detection of pig key points. Both the YOLOv8 and the VitPose models for detection of the human and her key points were initially downloaded from online repositories (MMPose Contributors., 2022; MMYOLO Contributors., 2022) as pre-trained models on the MS COCO dataset. The models were re-trained with 100 images with human skeletons labeled in the test arena. The YOLOv8 and the VitPose models for detection of the pig and its key points were trained on a dataset collected and labeled in the study of Juarez et al. (2023) merged with 100 images with pig skeletons labeled

in the test arena in the current study. As the MS COCO dataset did not contain a pig object class and the dataset collected and labeled in the study of Juarez et al. (2023) a human object class it was practical to use two separate models in the current study for detection of both classes of objects. All four models were trained for 1000 epochs with validation on the 50 labeled images on every 5th epoch. The progress in training of the models was evaluated with mean Average Precision (mAP) metric, whereas for the VitPose model Object Key Point similarity (OKS) measure was used (Lin et al., 2014) with per key point class standard deviations estimated in the study of Juarez et al. (2023). The 4 models with the highest mAP on the respective validation sets were selected for automated feature variables extraction on all video material recorded in the study.

2.8. Human-pig distance

The distance between the pig and the human was calculated according to Eq. 3,

$$D\left(a,b\right) = 0.59 * \sqrt{\left(a_x - b_x\right)^2 + \left(a_y - b_y\right)^2} \quad [cm] \tag{3}$$

where $D(a,b)$ is the Euclidean distance between pig's key point a, and human's key point b. The Cartesian coordinates of key body point a was denoted by $(a_x,\ a_y)$ and that of key body point b by $(b_x,\ b_y)$. The result of calculation of the Euclidean distance between key body points of the human and the pig expressed in pixels was multiplied by 0.59 to convert the distance to centimeters. Calculation of the distance was done on every frame of the dataset between all key body points of the human and the pig. The shortest distance between a pair of key points was selected as the human-pig distance feature variable (Fig. 4a).

2.9. Pig's visual attention proxy towards pen areas and the human

There were three areas inside the test arena important for the analysis of the pig's behavior in the context of human animal-interaction: (1) the window to the adjacent pen with companion pigs, (2) the entrance to

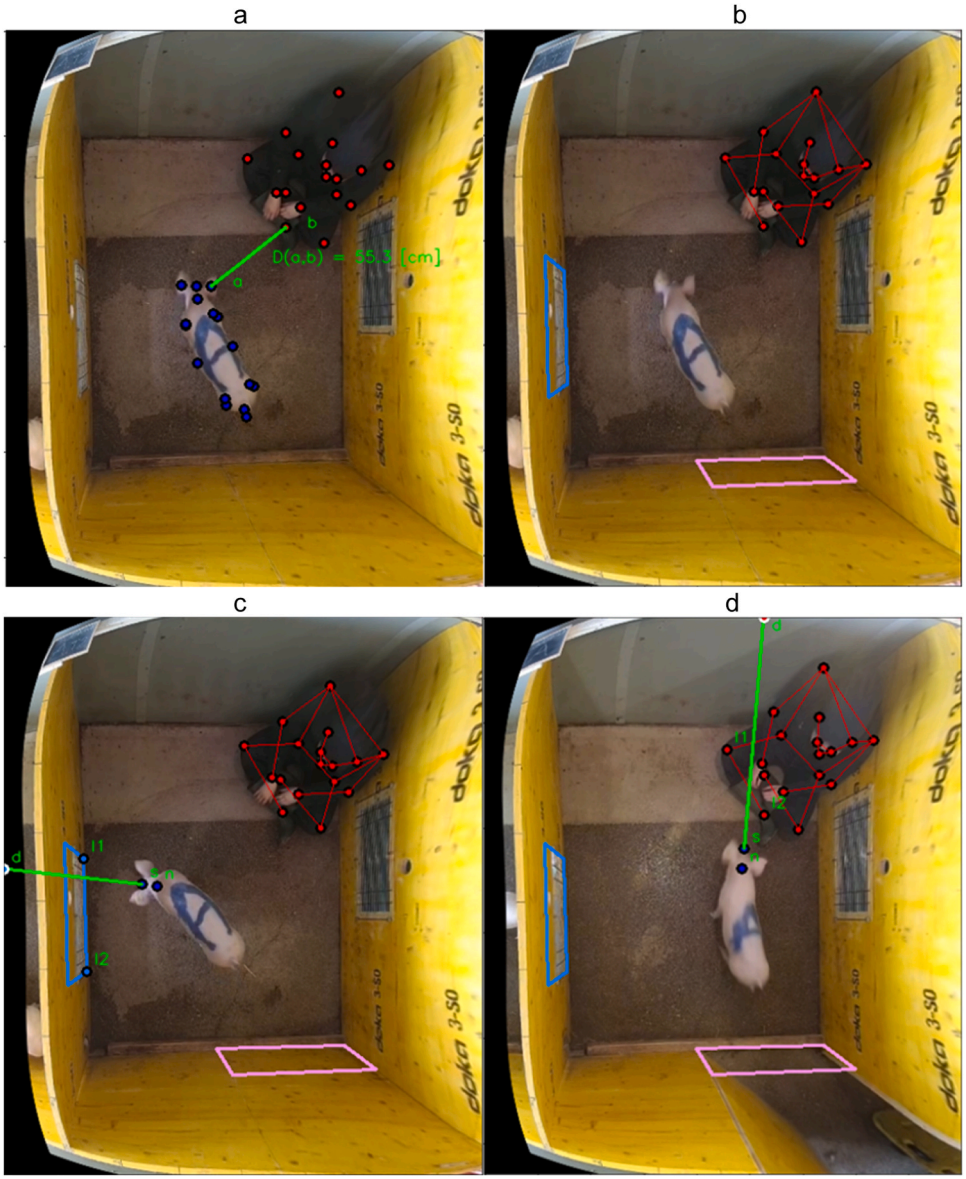


Fig. 4. Typical feature variables for research on human-animal interactions: a) Distance between the closest key points of a pig and a human, i.e. right ear and right ankle. b) Three areas in the test arena: the window to the companion arena in blue color, the entrance to the arena in pink color and the human skeleton in red color. c) Pig's visual attention proxy directed at the window. d) Pig's visual attention proxy directed at the human.

the arena, and (3) the human. The first two areas, which were located in the same position for the duration of the experiment, were specified as two quadrangles with fixed coordinates, while the human, whose position in the test arena was variable, was specified by automatically detected key points and their connections (Fig. 4b). Note that we estimated pig's visual attention proxy based on a line between the neck and the snout of the pig, but the pig has a large field of panoramic vision in addition to their front binocular vision (Prince, 1977). Therefore, this is only an approximation of the direction towards which they may be paying attention as pigs can also see sideways.

The first step in the process of detection of the pig's visual attention proxy towards the three areas in the test arena was to calculate a slope of a line between the neck and the snout of the pig according to Eq. 4,

$$slope = \frac{n_y - S_y}{n_x - S_x} \tag{4}$$

where (n_x, n_y) denoted the Cartesian coordinates of key point neck, while (s_x, s_y) denoted the Cartesian coordinates of key point snout. To detect the forward direction of the pig's head we firstly calculated the coordinates of 2 points at the end of the line between the neck and the snout of the pig according to Eqs. 5, 6, 7 and 8,

$$e_x = n_x + len \tag{5}$$

$$e_y = slope(e_x - n_x) + n_y \tag{6}$$

$$f_x = s_x - len \tag{7}$$

$$f_y = slope(f_x - s_x) + s_y \tag{8}$$

where (e_x, e_y) denoted the Cartesian coordinates of the first of the two points at the end of the line, while (f_x, f_y) denoted the Cartesian coordinates of the second of the two points at the end of the line. To ensure that the projected line between the neck and the snout of the pig spanned across the whole image reaching all the three areas in the test arena, we either added or subtracted a constant *len* with the value of 1000 pixels from the coordinates of the snout and the neck. In the following step Euclidean distances between the snout and the two points at the end of the line were calculated according to Eqs. 9 and 10,

$$D(s,e) = \sqrt{(s_x - e_x)^2 + (s_y - e_y)^2} \tag{9}$$

$$D(s,f) = \sqrt{(s_x - f_x)^2 + (s_y - f_y)^2} \tag{10}$$

where $d(s,e)$ is the Euclidean distance between the pig's key point snout and the first of the two points at the end of the line, while $d(s,f)$ is the Euclidean distance between the pig's key point snout and the second of the two points at the end of the line. The shorter of the 2 Euclidean distances indicated the pig's head direction (Figs. 4c and 4d).

On the basis of a line indicating the direction of the pig's head it was possible to estimate the attention of the pig towards the three areas inside the arena: the window to the companion arena with pigs, the entrance to the arena and the human. The direction of the pig's head was calculated by checking if there was an intersection of the line indicating the pig's head direction and (1) any of the 4 lines of the quadrangle drawn around the window to the adjacent pen or (2) any of the 4 lines of the quadrangle drawn around the entrance to the arena or (3) any of the 17 lines between the key points of the human' skeleton (Figs. 4c and 4d). The intersection between the line indicating the pig's head direction and any of the lines specifying the three areas was calculated according to Eqs. 11 and 12,

$$I1 = [(l2_y - l1_y)(l2_x - l1_x) - (l2_x - l1_x)(l2_y - l1_y)]\left[(l2_y - l1_y)(l2_x - l2_x) - (l2_x - l1_x)(l2_y - l2_y)\right] \tag{11}$$

$$I2 = \left[(d_y - s_y)\ (d_x - l1_x) - (d_x - s_x)(d_y - l1_y)\right]\left[(d_y - s_y)\ (d_x - l2_x) - (d_x - s_x)\ (d_y - l2_y)\right] \tag{12}$$

where $I1$ and $I2$ denoted two cross products of the 2 lines. The Cartesian coordinates of the snout was denoted by (s_x, s_y), of the point at the end of the line indicating the direction of the pig's head by (d_x, d_y). The Cartesian coordinates of any of the lines specifying the 3 areas inside the arena was denoted by $(l1_x, l1_y)$ and $(l2_x, l2_y)$. The intersection between both lines was registered if both $I1$ and $I2$ had values lower than 0. If either $I1$ or $I2$ was higher than 0 the pig was not directed to any of the three designated areas in the test arena.

3. Results

The YOLOv8 models trained to detect the human and the pig in the test arena reached similar performance of 0.86 mAP and 0.85 mAP, respectively on 55th and 75th epochs (Figs. 5a and 5b). The performance was similar despite the fact that the model for the detection of the human was trained on 250,073 human instances with 250,000 instances from COCO dataset and 73 from the test arena, whereas the model for detection of the pig was trained on 5083 pig instances with 5016 instances from the dataset of Juarez et al. (2023) and 67 from the test arena. In contrast to the object detection model, the VitPose model for skeleton detection of the human reached lower performance of 0.65 mAP on the 985th epoch than the VitPopse model for skeleton detection of the pig that reached 0.78 mAP on the 440th epoch (Figs. 5c and 5d).

The average distance between the human and the pig in the test arena was 31.03 cm (SD = 35.99). Out of the total duration of human-pig interaction tests of 35 h, pigs spent half of the time (17.5 h) within 10.44 cm distance to the human or closer (Fig. 6a). This was derived from the median distance between the human and the pig.

Out of the three predefined areas in the arena i.e. the window to the companion arena with pigs, the entrance to the arena and the human, pigs spend the most time with their head directed at the human i.e. 12 hrs 11 min (34.83 %). Pigs' heads were directed towards the entrance or the window for a similar duration of 3 hr 22 min (9.66 %) and 4 hrs (11.47 %) respectively (Fig. 6b).

4. Discussion

The better performance of the model for detection of the pig's skeleton than the model's for detection of the human's skeleton (0.78 mAP vs 0.65 mAP) might be related to the fact that human instances in the COCO dataset are mostly recorded from the frontal perspective or from a slight angle view, while the human in the test arena were only recorded in top view in our study dataset. Thus, it is possible that the pre-trained ViTPose model needed more examples with human skeletons in the test arena, in top view to reach similar performance to the VitPose model that was trained on the pig images. Pigs were recorded mostly in top view in both data subsets used for training i.e. in the dataset of Juarez et al. (2023) and from the current study in the test arena, hence offering a more homogeneous dataset. Similar performance of the YOLOv8 model trained to detect the human (0.86 mAP) and YOLOv8 model trained to detect the pig (0.85 mAP) suggests a better robustness of YOLOv8 to changes in camera perspectives than of the ViTPose model, or more limited possibility to re-train a large pre-trained ViTPose model to a different camera perspective. Further investigation and tests are required with other skeleton detection models than the ViTPose model,

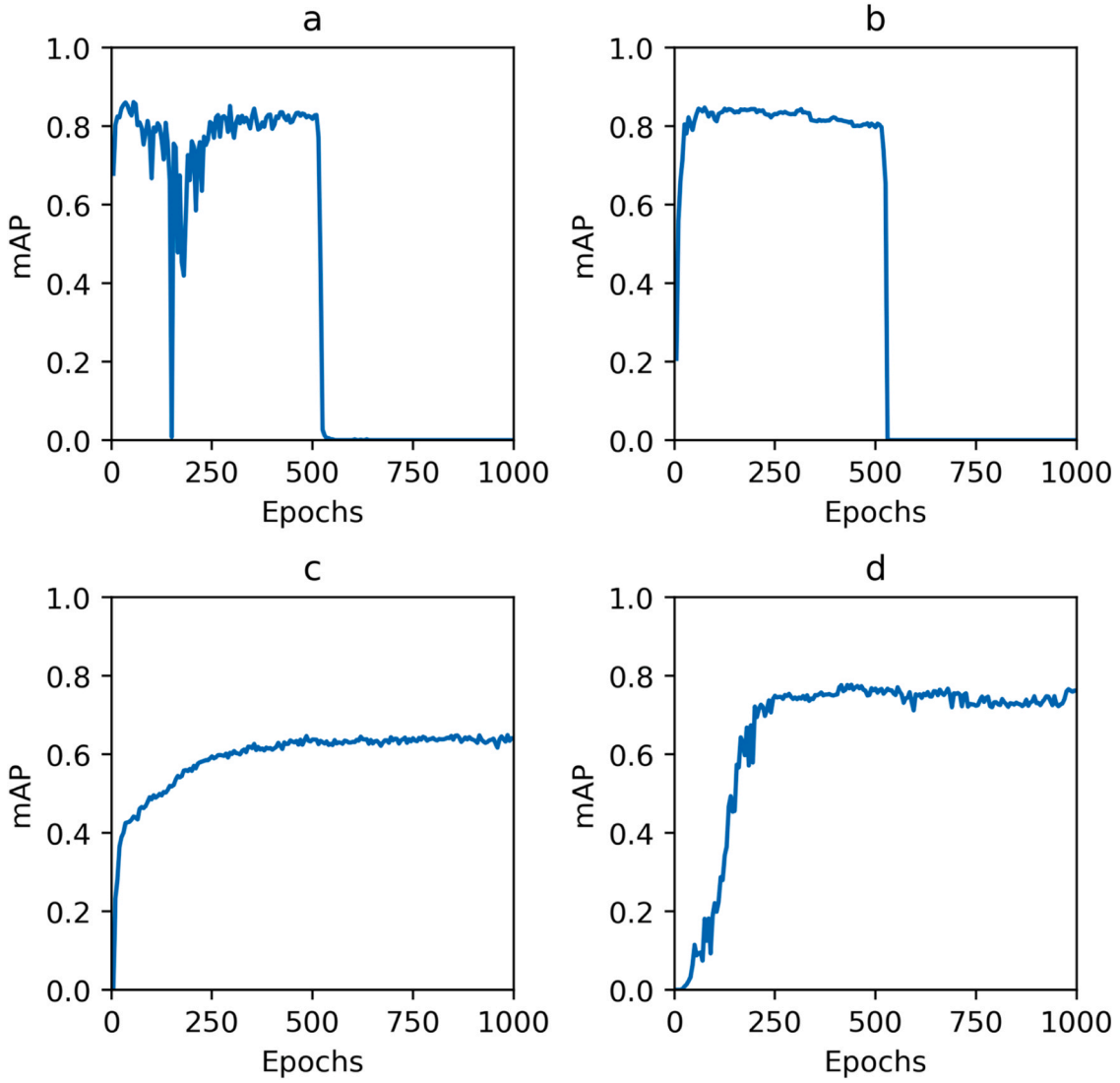


Fig. 5. Results of validation of the YOLOv8 and ViTPose models on every 5th epoch during the training process a) The YOLOv8 model trained to detect the human. B) The YOLOv8 model trained to detect the pig. C) The ViTPose model trained to detect human's 17 key point skeleton. D) The ViTPose model trained to detect pig's 16 key point skeleton.

e.g. RSN (Cai et al., 2020) to verify if this phenomenon is model-specific.

The results of the ViTPose model trained for detection of pig's skeleton with 0.78 mAP were similar to results obtained in the study of Juarez et al. (2023) in which the ViTPose model achieved a performance of 0.82 mAP. Slightly better performance of the model in the study of Juarez et al. (2023) might be related to the presence of more pigs from the same environment in the training and the validation sets, i.e. 585 instances in the training set in the study of Juarez et al. (2023) versus 67 instances in the current study. A similar paradigm of probable influence of dataset size on model performance was observed when comparing the performance of object detection model YOLOv8 for the detection of pigs with the performance of object detection model YOLOX applied in the study of Oczak et al. (2023) a training dataset with 9969 instances versus 67 instances in the current study. YOLOX achieved a performance of 96.5 mAP in the study of Oczak et al. (2023) in comparison to 0.85 mAP of YOLOv8 in the current study. Thus, YOLOX had better performance than YOLOv8 despite the fact that YOLOX is an older model achieving worse performance than YOLOv8 on the MS COCO benchmark dataset with 80 object categories.

When visually exploring the results of object and key point detection models no misclassification was observed that could negatively affect the measurement of head orientation or distance between the pig and the human. This performance was achieved despite the fact that few images from the test arena were used to either re-train or train the models i.e. 73 human instances and 67 pig instances. This is promising as in the future applications, e.g. for studies on human-animal relationships, time will be saved on laborious labeling of dataset specific pig's and human's skeletons.

The average distance between the human and the pig detected in the present study (31.03 cm) was shorter than the average distance of 56–85 cm in the "walking human test" in the study of Tanida et al. (1995) in which the human walked inside the arena after initial contact with the pig, although their area was 64 % larger (240 cm×160 cm) than in the present study (147 ×168 cm).

In the study of Hemsworth et al. (1986) in an arena test of 3 min duration pigs in three study groups spent on average 58 s (32.2 % of test duration), 47 s (26.1 %) and 26 s (14.4 %) in a proximity of 0.5 m to a human. The three groups of piglets received various handling by humans before the arena tests. The first group received regular handling (i.e. 3 min daily) by humans from 12 hrs after birth until 8 weeks of age. The second group received regular handling (i.e. 3 min daily) by humans from 27 days after birth until 8 weeks of age. The last group got routine

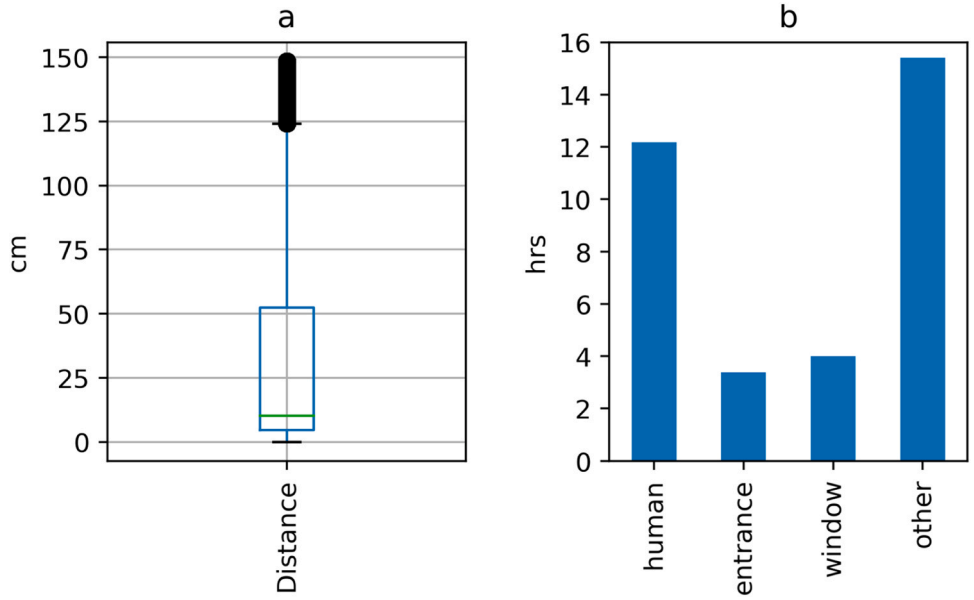


Fig. 6. Feature variables a) Distance between the pig and the human. B) Pig's visual attention proxy towards predefined areas in the test arena.

husbandry handling by humans. The average time spent in the proximity of 0.5 m to the human in our study was 7 min 13 s (72.3 %) in a 10 min test, which was more than for any of the groups in the study of Hemsworth et al. (1986) percentage wise. Pigs in our study were habituated to the presence of a human for 10 min daily in a period from about 28 days to 8 weeks of age, so very similar to the second group in the study of Hemsworth et al. (1986). Large proportion of time spent in close proximity to a human in our study might confirm the results of Hemsworth et al. (1986) indicating that habituation to the presence of a human in early life influences the behavioral response of pigs to humans later in life. However, the much smaller size of the arena in our study, which was nearly six times smaller (2.5 m^2 vs 14.4 m^2) than in the study of Hemsworth et al. (1986), possibly influenced the proportion of time spent in close proximity to a human.

Interestingly the time spent looking at the human in our study was longer percentage wise (34.83 %) than in the result obtained by Tallet et al. (2014) for pigs that received gentle tactile contact from a human in a period from day after weaning until approx. 8 weeks of age i.e. on average 8 s (6.7 %) and 17.5 s (14.6 %) in a 2 min test. The reason for this difference, as in the studies cited above for comparison of time spent in close proximity to a human, might be the bigger size of the test arena, which was approximately three and a half times bigger (2.5 m^2 vs 9 m^2) in the study of Tallet et al. (2014) than in our study.

A similar methodology for the estimation of the pig's head direction to the one applied in the current study was used in the study of Oczak et al. (2022). In this study the sow's head direction was estimated based on a line projected between the nose and the point between the ears of the sow instead of the neck as in our study. The distance between the closest point on this line and the center of the hay rack indicated the degree to which the sow's head was directed towards the hay rack in the pen. This feature variable was then used together with several other feature variables to estimate the time spent by the sow on the use of hay rack in the pen. In our current study the estimated head's direction was used to binary classify attention of the pig towards predefined areas in the pen by estimation of intersection with lines specifying the location of the areas i.e. attention was directed towards the area or not rather than attention expressed as continuous variable as in the study of Oczak et al. (2022). Binary classification of attention of the pig towards pen areas based on intersecting lines has the advantage of easier interpretability without the need of introducing additional thresholding on the calculated variable similarly as in the study of Oczak et al. (2022). Another difference between both studies was the type of models used to detect the body parts. In the study of Oczak et al. (2022) both the ears and the nose were detected with object detection model RetinaNet instead of the key body point detection model as in the current study. This potentially reduces the robustness of detections towards occlusions in the pen as skeleton detection models should be more robust to occlusions, deformations and novel poses (Newell et al., 2016; Toshev and Szegedy, 2013). Note however that as we cautioned before, pigs have a larger field of vision than for example humans and other predator species, combining a large panoramic vision on their sides and a narrower binocular vision in front of them, and therefore the direction of their head is only an approximation of their visual attention (Prince, 1977).

We successfully developed a method based on computer vision to automatically measure the distance and visual attention that could be directly applied in human-animal relationship tests. This approach could also be used in other types of behavioral tests such as open field and novel object tests to measure the animal's reaction towards objects or features of their environment. With slight modifications the method could be used in human-animal relationship tests to measure other aspects of human-animal interaction such as approaching, following or avoiding the human, object, situation or other variables of interest (for a review, see Waiblinger et al., 2006 or Rault et al., 2020). The method could be used in a scoring system of the test, such as pig withdraws at a distance of $>$ one arm's length, pig withdraws at a distance of $\leq$ one arm's length, pig can be touched at the snout or at the head, or pig can be touched behind the at the snout, head, ears and neck.

The first out of two examples of how the method could be simply adapted to extend its application might be in novel object tests to measure touching the object. The object would have to be detected first with an object detection model. Then it would be necessary to decide on a distance around pig's key points that indicate the touch e.g. 3 cm around the snout. The second example of the adaptations of the method, this time to detect interaction with the human, could involve deciding on specific human key points that need to be in close proximity to pig key points, as probably not all human key points are relevant for the interaction e.g. nose or shoulders. Then deciding on a threshold distance between pig's key points and human key points would be needed e.g. wrists in 3 cm proximity to any pig's key point.

In order to extend the method to behavioral tests involving groups of animals such as latency until the first pig of the group approaches the novel object or the human, the method of tracking individuals in a group

would have to be integrated with the method presented in the current study. For example, tracking methods based on Graph Convolutional Networks (Parmiggiani et al., 2023) could be used. Alternatively, object detection methods, which simply detect distinctive spray marks on the back of the pigs usually used in behavioral tests, could be used.

The attitude and behavior of stockpeople affect the animals' fear of humans, which ultimately influence animals' productivity and welfare (Zulkifli, 2013). Thus, a very interesting area of application of the method is for automatic detection of human-animal interactions in the practical farm environment to evaluate the quality of care by staff on farms. This could be integrated with an evaluation of farm animal welfare such as the Welfare Quality Protocol (Welfare Quality®., 2009) or the implementation of routine monitoring with computer vision.

5. Conclusions

The developed computer vision method to automatically measure distance and visual attention proxy could be directly applied in human-animal relationship tests. This approach could also be used in other types of behavioral tests such as open field and novel object tests to measure the animal's reaction towards objects or features of their environment. The applications of our method offer a possibility to extract those variables in an automated way, reducing the need for time-consuming manual observation or labeling of videos. It can also support objective analysis of animal behavior and human-animal relationships.

Ethical statement

All procedures involving animal handling and treatment were approved by the institutional ethics committee of the University of Veterinary Medicine Vienna and the National authority according to the Law for Animal Experiments, Tierversuchsgesetz (GZ 2022 – 0.118.102).

CRediT authorship contribution statement

Maciej Oczak: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **Jean-Loup Rault:** Writing – review & editing, Resources, Funding acquisition. **Suzanne Truong:** Writing – review & editing, Investigation. **Oceane Schmitt:** Writing – review & editing, Supervision, Project administration, Investigation, Data curation, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank the farm staff and technicians who took part in data collection. The animal experimental part of this research was funded through a Austrian FWF project grant # P 33669-B n.

References

Bensoussan, S., Tigeot, R., Meunier-Salaün, M.-C., Tallet, C., 2020. Broadcasting human voice to piglets (Sus scrofa domestica) modifies their behavioural reaction to human presence in the home pen and in arena tests. Appl. Anim. Behav. Sci. 225, 104965.

Bradski, G., 2000. The openCV library. Dr. Dobb's. J.: Softw. Tools Prof. Program. 25, 120–123.

Brooks, J., 2019. COCO Annotator. URL https://github.com/jsbroks/coco-annotator (accessed 12.1.23).

Cai, Y., Wang, Z., Luo, Z., Yin, B., Du, A., Wang, H., Zhang, X., Zhou, X., Zhou, E., Sun, J., 2020. Learning Delicate Local Representations for Multi-person Pose Estimation, in: Computer Vision – ECCV 2020. Springer International Publishing, pp. 455–472.

Czycholl, I., Menke, S., Straßburg, C., Krieter, J., 2019. Reliability of different behavioural tests for growing pigs on-farm. Appl. Anim. Behav. Sci. 213, 65–73.

Forkman, B., Boissy, A., Meunier-Salaün, M.-C., Canali, E., Jones, R.B., 2007. A critical review of fear tests used on cattle, pigs, sheep, poultry and horses. Physiol. Behav. 92, 340–374.

Grabovskaya, S.V., Salyha, Y.T., 2014. Do results of the open field test depend on the arena shape? Neurophysiology 46, 376–380.

Hemsworth, P.H., Barnett, J.L., Hansen, C., Gonyou, H.W., 1986. The influence of early contact with humans on subsequent behavioural response of pigs to humans. Appl. Anim. Behav. Sci. 15, 55–63.

Juarez, S., Kielar, A., Drabik, A., Stec, A., Stós-Wyżga, Z., Nowicki, J., Oczak, M., 2023. Standardisation of the Structure of Pig's Skeleton for Automated Vision Tasks. https://doi.org/10.2139/ssrn.4659489.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: Common Objects in Context, in: Computer Vision – ECCV 2014. Springer International Publishing, pp. 740–755.

Ling, Y., Jimin, Z., Caixing, L., Xuhong, T., Sumin, Z., 2022. Point cloud-based pig body size measurement featured by standard and non-standard postures. Comput. Electron. Agric. 199, 107135.

MMDetection Contributors, 2018. OpenMMLab Detection Toolbox and Benchmark. URL https://github.com/open-mmlab/mmdetection (accessed 12.1.23).

MMPose Contributors, 2022. OpenMMLab Pose Estimation Toolbox and Benchmark [WWW Document]. OpenMMLab Pose Estimation Toolbox and Benchmark. URL ⟨https://mmpose.readthedocs.io/en/latest/model_zoo_papers/algorithms.html#vitpose-neurips-2022⟩ (accessed 9.1.23).

MMYOLO Contributors, 2022. MMYOLO: OpenMMLab YOLO series toolbox and benchmark [WWW Document]. MMYOLO: OpenMMLab YOLO series toolbox and benchmark. URL ⟨https://github.com/open-mmlab/mmyolo/tree/main/configs/yolov8⟩ (accessed 9.1.23).

Newell, A., Yang, K., Deng, J., 2016. Stacked Hourglass Networks for Human Pose Estimation. In: Computer Vision – ECCV 2016. Springer International Publishing, pp. 483–499.

Oczak, M., Bayer, F., Vetter, S.G., Maschat, K., Baumgartner, J., 2022. Where Is Sow's Nose?-RetinaNet Object Detector As A Basis For Monitoring Use Of Rack With Nest-Building Material. Front. Anim. Sci. 3, 92.

Oczak, M., Maschat, K., Baumgartner, J., 2023. Implementation of Computer-Vision-Based Farrowing Prediction in Pens with Temporary Sow Confinement. Vet. Sci. China 10. https://doi.org/10.3390/vetsci10020109.

Parmiggiani, A., Liu, D., Psota, E., Fitzgerald, R., Norton, T., 2023. Don't get lost in the crowd: Graph convolutional network for online animal tracking in dense groups. Comput. Electron. Agric. 212, 108038.

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. Nat. Methods 16, 117–125.

Prince, J.H., 1977. The eye and vision. In: Swenson, M.J. (Ed.), Dukes Physiology of Domestic Animals. Cornell University Press, New York, pp. 696–712.

Rault, J.-L., Waiblinger, S., Boivin, X., Hemsworth, P., 2020. The Power of a Positive Human-Animal Relationship for Animal Welfare. Front Vet. Sci. 7, 590867.

Solawetz, J., 2023. What is YOLOv8? The Ultimate Guide [WWW Document]. Roboflow Blog. URL ⟨https://blog.roboflow.com/whats-new-in-yolov8/⟩ (accessed 12.17.23).

Tallet, C., Sy, K., Prunier, A., Nowak, R., Boissy, A., Boivin, X., 2014. Behavioural and physiological reactions of piglets to gentle tactile interactions vary according to their previous experience with humans. Livest. Sci. 167, 331–341.

Tanida, H., Miura, A., Tanaka, T., Yoshimoto, T., 1995. Behavioral response to humans in individually handled weanling pigs. Appl. Anim. Behav. Sci. 42, 249–259.

Toshev, A., Szegedy, C., 2013. DeepPose: Human Pose Estimation via Deep Neural Networks. arXiv [cs.CV].

Waiblinger, S., Boivin, X., Pedersen, V., Tosi, M.-V., Janczak, A.M., Visser, E.K., Jones, R. B., 2006. Assessing the human–animal relationship in farmed species: A critical review. Appl. Anim. Behav. Sci. 101, 185–242.

Wang, X., Wang, W., Lu, J., Wang, H., 2022. HRST: An Improved HRNet for Detecting Joint Points of Pigs. Sensors 22. https://doi.org/10.3390/s22197215.

Wang, Z., Zhou, S., Yin, P., Xu, A., Ye, J., 2023. GANPose: Pose estimation of grouped pigs using a generative adversarial network. Comput. Electron. Agric. 212, 108119.

Welfare Quality® 2009. Welfare Quality® assessment protocol for pigs (sows and piglets, growing and finishing pigs). Welfare Quality® Consortium, Lelystad, Netherlands.

Xu, Y., Zhang, J., Zhang, Q., Tao, D., 2022. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. arXiv [cs.CV].

Zulkifli, I., 2013. Review of human-animal interactions and their impact on animal productivity and welfare. J. Anim. Sci. Biotechnol. 4, 25.