Smart Agricultural Technology 14 (2026) 102303

Contents lists available at ScienceDirect

Smart Agricultural Technology

journal homepage: www.journals.elsevier.com/smart-agricultural-technology

Pose estimation in pigs: Comparative analysis of keypoint configurations and neural network architectures

Kirill Ivanov [a], Valentina Bonfatti [a], Claudia Kasper [b], Hassan-Roland Nasser [c,*]

[a] Department of Comparative Biomedicine and Food Science, University of Padova, Viale dell'Università 16, 35020, Legnaro (PD), Italy
[b] Animal GenoPhenomics, Agroscope, Posieux, Switzerland
[c] Digital Production, Agroscope, Posieux, Switzerland

ARTICLE INFO

Keywords:
Pose estimation
Computer vision
Precision livestock farming
Animal monitoring
Keypoints

ABSTRACT

Accurate and robust pose estimation is essential for automated livestock monitoring, particularly in group-housed pigs where occlusions, animal overlap, and variable orientations complicate keypoint detection. This study evaluated the effects of keypoint configuration and neural network architecture on pose estimation performance in group-housed fattening pigs. Five skeletal configurations were first compared using ResNet-50 in DeepLabCut: a standard 7-keypoint model and four extended configurations including additional ear, dorsal body, eye, or tail-related landmarks. BODY5P, which included five anatomically distributed landmarks along the dorsal body axis, achieved the best overall test performance, with the highest mAP (75.12 ± 1.42%) and the lowest RMSE (7.43 ± 0.43 px), suggesting that consistently visible dorsal landmarks improve robustness under group-housing conditions. Five neural network architectures were then compared using a larger annotated dataset and repeated train-test evaluation. HRNet-W32 achieved the highest accuracy (test mAP: 92.52%; RMSE: 7.93 ± 0.36 px), whereas lighter DLCRNet architectures trained faster and showed competitive test performance, indicating their potential for resource-constrained or frequently retrained on-farm applications. Temporal analysis of high-confidence predictions from an independent 15-min video showed stable tracking for most dorsal landmarks, while the snout had the highest dropout rate (~46%). Exploratory spatial analysis showed that pose outputs can support occupancy mapping and tail-snout proximity analysis, although proximity events should be interpreted as candidate interaction indicators rather than confirmed tail-biting events. Overall, these results highlight the importance of balancing anatomical relevance, landmark visibility, model accuracy, and computational efficiency when designing pose-estimation pipelines for pig behaviour monitoring.

1. Introduction

Monitoring animal behaviour is particularly important in pig production, where aggression and tail biting remain major welfare and management challenges. These behaviours can compromise animal welfare and lead to economic losses through increased veterinary costs, reduced growth performance, elevated mortality, and carcass downgrading (Devi, S. J., et al., 2024; [1]). Continuous monitoring of pigs is therefore essential for two complementary purposes: first, to improve the understanding of the behavioural and environmental factors that contribute to welfare problems, and second, to support early detection and timely intervention before severe damage occurs [2].

Recent advances in computer vision and machine learning have created new opportunities for automated and non-invasive monitoring of livestock behaviour. Video-based methods can be used to detect animals, track their movement, identify postures, and recognize complex behaviours such as aggression, social interaction, or mating-related activity [1,3]. Compared with manual observation, automated video analysis can provide continuous and high-resolution behavioural information while reducing observer workload and subjectivity. These advantages make computer vision particularly relevant for practical farm environments, where long-term and scalable monitoring is required.

Deep learning-based pose estimation algorithms, initially developed for human image analysis, have become essential tools for identifying and tracking animal behaviour (Jayachitra [4]). These models are capable of managing substantial variability in body morphology, environmental conditions, and image quality by leveraging large, annotated datasets during training. While behavioural studies in laboratory

* Corresponding author at: Digital Production, Agroscope, Rte de la Tioleyre 4, 1725, Posieux, Switzerland.
E-mail address: hassan-roland.nasser@agroscope.admin.ch (H.-R. Nasser).

https://doi.org/10.1016/j.atech.2026.102303
Received 22 January 2026; Received in revised form 11 May 2026; Accepted 12 June 2026
Available online 13 June 2026

settings typically benefit from controlled environments, their imaging conditions are often tightly constrained by the specific experimental design, limiting generalizability. To address this challenge, recent approaches have adopted transfer learning techniques, in which neural networks pre-trained on general object classification tasks are fine-tuned on smaller, domain-specific datasets of animal images [5]. In this study, pose estimation refers to the detection of predefined anatomical landmarks, or keypoints, on the pig body and their connection into a skeletal representation. This representation does not directly classify behaviour but provides spatial and temporal body-part information that can be used for downstream behavioural monitoring.

However, pose estimation in group-housed pigs remains challenging. Animals frequently overlap, occlude one another, change orientation, and interact in close proximity. These factors complicate both keypoint localization and the association of detected keypoints with individual animals across time. In addition, anatomical landmarks differ in their visibility and behavioural relevance. Dorsal body landmarks may be easier to detect because they are relatively large and consistently visible from overhead views, whereas facial or tail-related landmarks may be more difficult to localize due to occlusion, motion, and posture variation. At the same time, these more challenging landmarks may be essential for specific behavioural applications, including the analysis of tail-directed interactions.

Previous studies have demonstrated the potential of deep learning for pig detection, posture classification, and pose estimation under farm or semi-farm conditions [6,7,1]. Recent work has also benchmarked several pose estimation frameworks, including YOLOv8-pose, AlphaPose, and OpenPose, for multi-object keypoint detection in group-housed pigs, with YOLOv8-pose showing strong performance in terms of accuracy and efficiency [7]. Nevertheless, several methodological questions remain insufficiently explored. First, it is not yet clear how the choice of keypoint configuration affects pose estimation performance in group-housed pigs, especially when balancing robust global body tracking against the need to include behaviourally specific landmarks such as the snout or tail. Second, direct comparisons among neural network architectures under the same experimental conditions are still limited. Third, the practical value of pose estimation for downstream behavioural analysis depends not only on overall localization accuracy, but also on the reliability of specific keypoints, temporal consistency of trajectories, and the interpretability of spatial interaction metrics. Therefore, the aim of this study was to evaluate the performance of alternative keypoint configurations and neural network architectures for pose estimation in group-housed fattening pigs. Specifically, we compared five skeletal configurations, including a standard 7-keypoint model and four extended configurations incorporating additional ear, dorsal body, eye, or tail-related landmarks. We then compared five neural network architectures implemented in DeepLabCut, namely HRNet-W32, DLCRNet S32, DLCRNet S16, ResNet-50, and ResNet-101. The analysis further explored whether pose-estimation outputs could support temporal tracking assessment, occupancy mapping, and candidate tail-snout proximity analysis as preliminary indicators for downstream behavioural monitoring.

2. Materials and methods

2.1. Experimental setting

The study was conducted at the Agroscope experimental pig facility (Posieux, Switzerland) and involved 12 castrated Large White male pigs, all born on-site at the Agroscope experimental farm and individually identified using RFID ear tags. The animals had intact tails, in compliance with Article 18(a) of the Swiss Animal Protection Ordinance. Observations began when the pigs reached approximately 100 kg live weight and 140 days of age. The pigs were housed in a single pen (Fig. 1) covering a total area of 32.14 m^2, which included a resting area with a solid floor (19.12 m^2), and an elimination area equipped with a slatted floor (13.02 m^2). They had ad libitum access to water and received dry feed through two automated single-spaced feeders (Schauer Maschinenfabrik GmbH & Co. KG, Prambachkirchen, Austria) operating between 07:30 and 23:30. The pen was divided into two contiguous interconnected zones of identical size and equipment. For each zone, enrichment and functional elements included one nipple drinker, one drinking bowl, two fixed and one suspended flexible straw basket, a feeding station, and two metal chains attached to the sidebars. Animals were allowed to move freely between the two zones at all times.

To enhance individual identification on video recordings, pigs were spray-marked weekly during routine weighing procedures practiced in fattening experiments, eliminating the need for additional direct handling. This method took advantage of the weighing process, during which pigs were individually separated and restrained, allowing for reliable ear tag verification and accurate reapplication of the markings. Regular re-marking ensured that the identification marks remained clearly visible throughout the video recording period. Blue spray dye was selected based on its high visual contrast and prolonged visibility under typical lighting and hygiene conditions. The marking procedure followed the standard approach of animals marking, consisting of drawing continuous lines across the pig's body from one side to the other, in different anatomical regions (i.e., rear, back and shoulders; Fig. 2). This technique proved effective for a small cohort and allowed for consistent identification regardless of the pigs' orientation or posture, as the markings were clearly visible from both lateral and overhead camera views.

These visual markings were used to support individual recognition during manual annotation and subsequent video inspection. Because

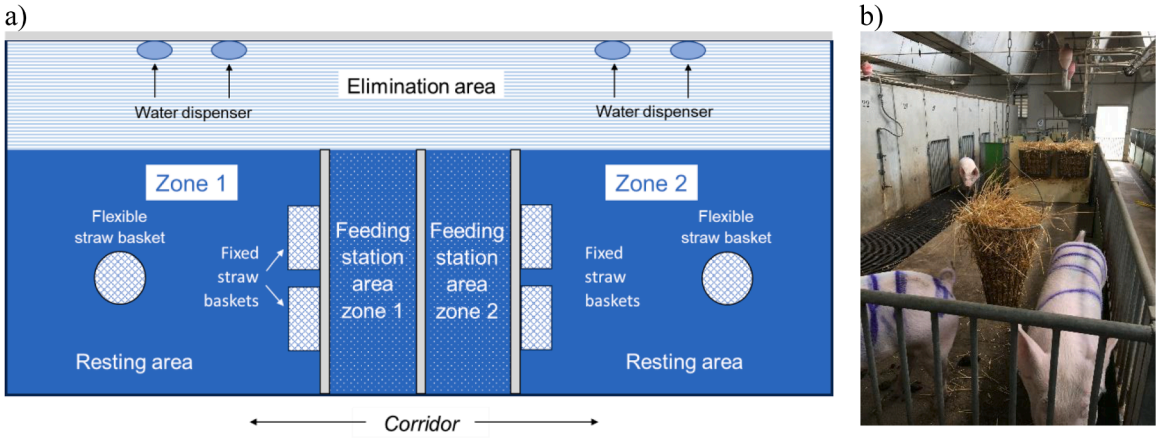

Fig. 1. Schematic representation (a) and picture (b) of the experimental pen.

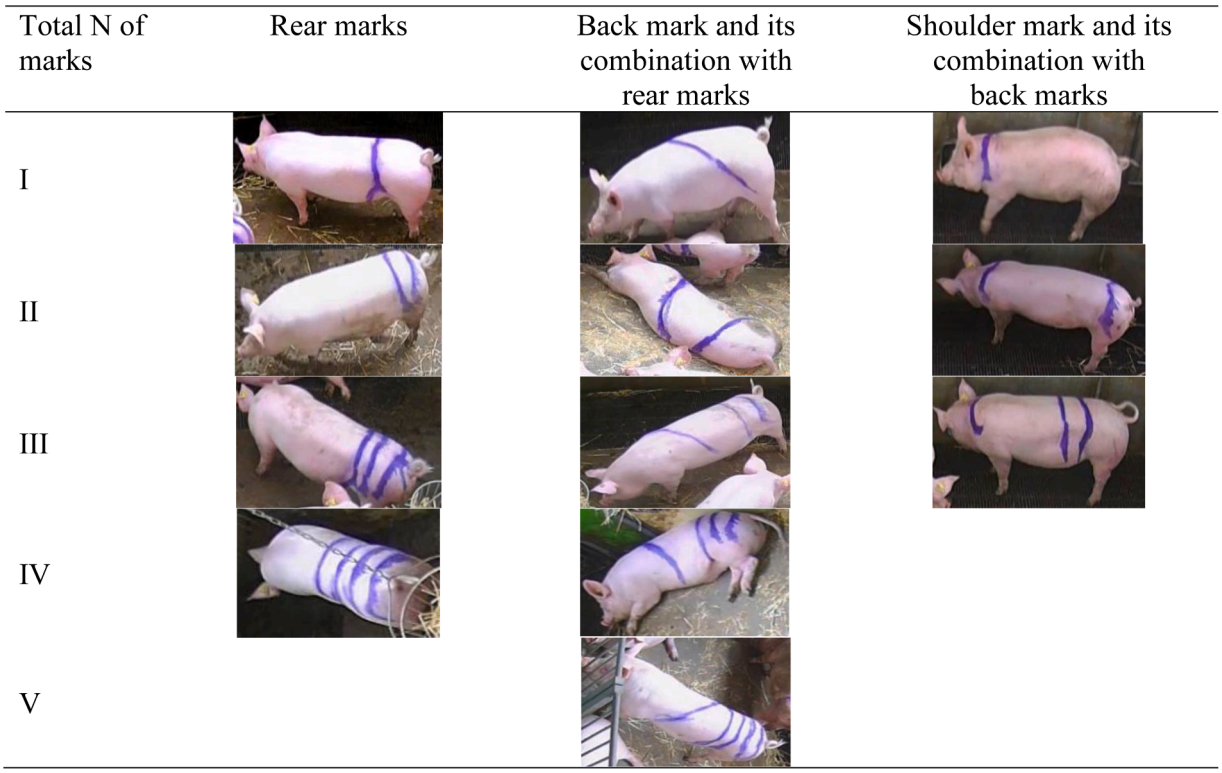

Fig. 2. Marks for individual pig identification on video recordings.

pigs were housed as a group and could overlap or occlude one another, individual identity was assigned only when the marking pattern and body position were visually clear. Cases in which identity assignment was uncertain due to overlap, occlusion, or poor visibility were treated conservatively and were not used for identity-dependent downstream analyses.

2.2. Video acquisition

Video recordings were obtained using 6 Panasonic I-Pro Mega Super Dynamic WV-SW316L (Panasonic Holdings Corporation, Kadoma – Osaka, Japan) surveillance cameras, which were installed above the pens to capture the different functional zones, including the resting areas and feeding stations. The cameras were installed at a height of approximately 2.5 m using pre-existing overhead supports and positioned at varying angles to the animals, with two cameras oriented at approximately 45° angles. This configuration was designed to maximize coverage of the region of interest while minimizing occlusions caused by animals, structural elements (e.g., walls, dividers, bars), and equipment. The camera system was connected to a QNAP Network-Attached Storage (**NAS**) device, providing 12 TB of digital storage capacity for video data. Video files were managed using VioStor System Network Access Storage management software (QNAP Systems, Inc., Xizhi District – New Taipei City, Taiwan), which enabled efficient playback and retrieval based on camera ID, date, and timestamp.

Video recordings were conducted daily, from 07:00 to 20:00, between April 25 and June 12, 2023, and covered both zones of the pen, under naturally varying lighting conditions influenced by the time of day and weather. In total, 12,415 video clips were collected, corresponding to approximately 3103 h of footage. However, due to a hardware malfunction in camera 6, the video files (n = 2 472) became corrupted and could not be retrieved.

To balance storage efficiency with sufficient visual quality for behavioural analysis, videos were recorded at a frame rate of 15 frames per second (fps) and a resolution of 320 × 240 px. Recordings were saved in AVI format, with each video lasting 15 min. To optimize data processing and reduce training time for convolutional neural networks (**CNNs**), the available 9943 AVI files were converted to MP4 format, which offers improved compression efficiency without significant loss of quality [8]. Following conversion, a video selection step was performed to exclude non-informative videos. Videos were removed if they met any of the following criteria: (i) no animals were present in the video or farm staff were visible inside the pens, primarily during morning cleaning (347 videos), or (ii) recordings corresponded to periods when pigs were not yet visibly marked for individual identification (April 25 to May 1 and June 8 to June 12, 2023; 2235 videos total).

Due to the design of the pen, each of the interconnected zones was monitored by a separate camera. Hence, the number of animals visible in each frame could vary. For body part recognition and pose estimation, videos recorded from camera 5 (Fig. 3) were selected for further analysis. For this camera, the dataset after editing consisted of 1834 15-min videos, each containing relevant video material with visible pigs, identifiable markings, and sufficient image quality for pose estimation analysis. This subset was used as the source material for frame extraction, annotation, model training, and downstream spatial behaviour analysis.

2.3. Keypoint annotation and comparison among different skeleton model configurations

To evaluate the effect of skeletal design on pose estimation performance, five different keypoint configurations were compared using frames extracted from the selected camera 5 video dataset. The annotated image set for this comparison was expanded to include 500 images, capturing variation in pig posture, orientation, overlap, and visibility under group-housing conditions. The same image set was used for all keypoint configurations to ensure that differences in performance reflected the anatomical landmark design rather than differences in image content. First, we defined a simplified skeletal model for pigs consisting of 7 anatomically relevant keypoints (standard keypoint configuration,

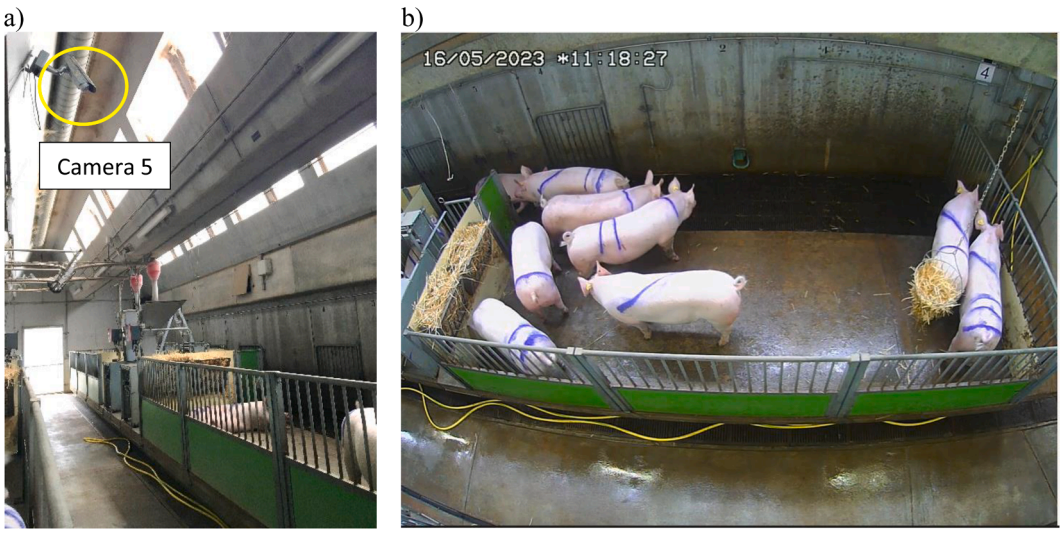

Fig. 3. Overhead camera setup showing positioning above the pen (a), and the corresponding recorded view of the pen area (b).

STD). The 7-keypoint STD configuration was selected as a baseline because it represented the minimal anatomical structure required to describe the main body axis and head-tail orientation of the pig while keeping annotation complexity low. This configuration provided a reference model against which the effect of adding more detailed anatomical landmarks could be evaluated.

As illustrated in Fig. 4, the model captures the overall structure of the animal using a series of custom-defined landmarks and linear connections forming a skeletal representation. The pose skeleton begins at the snout (Point 1), serving as the initial anchor point for skeleton construction. The head point (Point 2) is defined at the dorsal surface of the skull, located medially between the bases of the ears. The left and right ear points (Points 3 and 4, respectively) are defined at the distal tips of the ear triangle, located symmetrically on either side of the head. The dorsal midline of the pig is represented by 2 keypoints (Points 5 and 6) forming the primary axis of the skeleton. These two dorsal keypoints are located along the midline of the back: the anterior point at the vertical projection from the forelimb, and the posterior point at the vertical projection from the hindlimb. The tail base point (Point 7) is positioned at the anatomical base of the tail and is connected sequentially to the second dorsal point. Four additional configurations were derived from STD and included extra landmarks on the ears (**EAR3P**), dorsal body axis (**BODY5P**), tail region (**TAIL2P**), or head region, including the eyes (**EYES2P**). Specifically, in the BODY5P configuration, five dorsal body landmarks were distributed along the midline of the back from the anterior shoulder region to the posterior rump region. These landmarks are referred to as Body1 to Body5 throughout the manuscript. All keypoint and skeleton configurations were annotated on the same set of 500

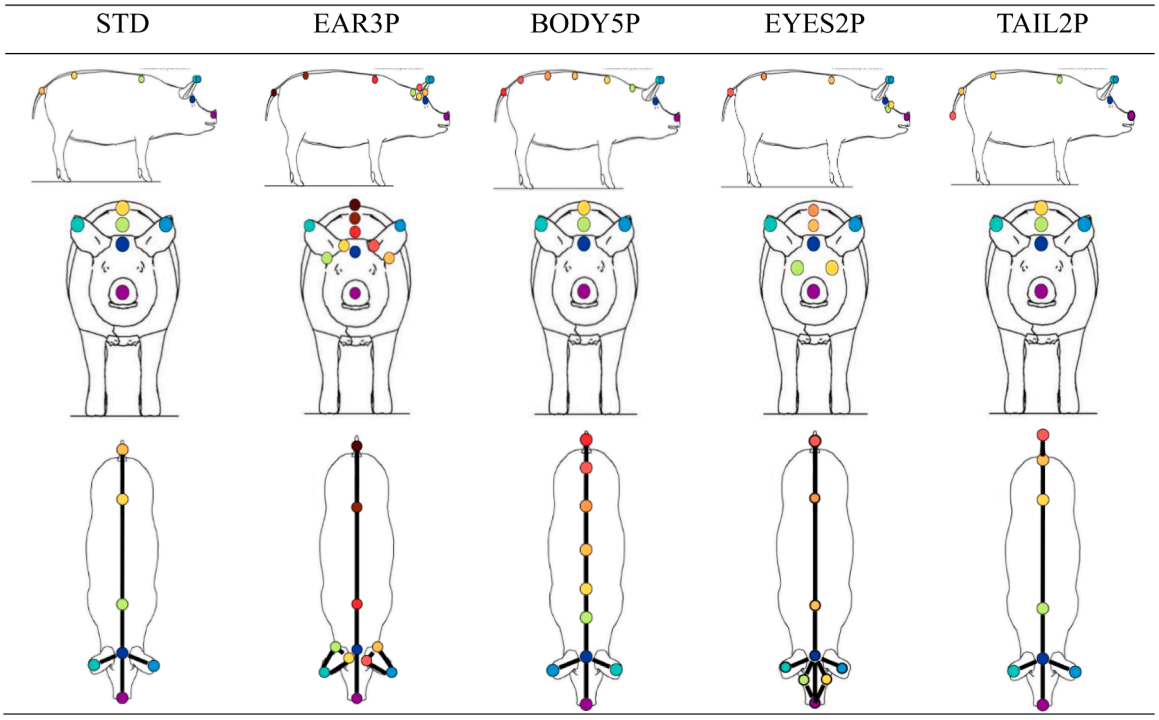

Fig. 4. Custom-defined landmarks and linear connections forming a skeletal representation across different keypoint configurations. The configurations included a 7-keypoint model (STD) and four variants with additional anatomical landmarks focused on the ears (EAR3P), body (BODY5P), eyes (EYES2P), and tail (TAIL2P).

images, ensuring consistency across comparative analyses. Anatomical keypoints were manually labelled using DeepLabCut (ver. 3.0.1; [3]). During annotation, each visible pig was treated as a separate animal instance when individual identity could be determined from the spray-marking pattern and spatial position within the frame. Keypoints were not assigned when the relevant body part was not visible or when the identity of the animal was ambiguous. The selection of image annotation software was guided by four main criteria: widespread adoption within the research community, availability of a user-friendly graphical interface (GUI), ease of use, and the availability of open-source solutions. DeepLabCut leverages deep learning techniques to accurately predict the coordinates of user-defined anatomical keypoints. Built on TensorFlow or PyTorch, it employs transfer learning, enabling pre-trained convolutional neural networks to be fine-tuned on relatively small datasets. These networks act as feature extractors, learning spatial and contextual information from images, which is then used to infer body part locations [3]. DeepLabCut's proven performance in dynamic and complex behavioural scenarios makes it particularly well-suited for the aim of the present study.

To compare the five keypoint configurations, pose estimation was performed using ResNet-50 as a common baseline architecture implemented in DeepLabCut. For each configuration, the annotated image set was divided into training and testing subsets using an 80:20 split, with 400 images used for training and 100 images withheld for testing. The same split was applied across all configurations to ensure a controlled comparison. This design allowed the effect of keypoint configuration to be evaluated independently of the neural network architecture and reduced the dependence of the comparison on a very small held-out test set.

2.4. Pose estimation and model training

After the comparison of keypoint configurations, a separate architecture-comparison experiment was conducted to evaluate pose estimation robustness and generalization across a wider range of video conditions. To build a diverse and representative dataset for this experiment, 10 videos were randomly selected before frame extraction. To capture the variability of pig postures and environmental conditions, a k-means clustering algorithm was applied individually to each video. This method grouped visually similar frames into clusters, enabling the automated selection of 100 frames per video that best represented the range of postures and positions present. Thus, the final pose estimation dataset comprised a total of 1000 frames, capturing the behaviour of 12 pigs at different times of day and under varying lighting conditions. These frames presented challenges typical of real farm environments, including non-uniform background colours, varying illumination intensities, and frequent occlusions or partial views of pigs.

The set of 1000 frames was used to evaluate pose estimation robustness and generalization across a wider range of conditions. While 1000 frames is a relatively small dataset for deep learning applications, this limitation is common in manual animal behaviour annotation due to practical constraints. To overcome this, transfer learning from pre-trained models was employed, which can improve model performance when the available annotated dataset is limited. Based on the keypoint configuration comparison, BODY5P was selected for the neural network architecture comparison because it provided the best balance between localization accuracy, anatomical coverage, and landmark visibility. This ensured that differences among models reflected differences in neural network architecture rather than differences in landmark definition.

Five neural network architectures, namely DLCRNet S16, DLCRNet S32, HRNet-W32, ResNet-50, and ResNet-101, implemented in DeepLabCut, were evaluated for their tracking accuracy, computational efficiency, and robustness. ResNet-50 and ResNet-101 are general-purpose convolutional neural networks with 50 and 101 layers, respectively. Both are widely used in computer vision tasks, including pose estimation, with ResNet-101 offering deeper feature extraction at the cost of slower performance and a higher risk of overfitting on small datasets. DLCRNet S16 and DLCRNet S32 are custom, truncated variants of ResNet-50 optimized for DeepLabCut; they are designed to reduce computational load while maintaining accuracy in animal pose estimation. HRNet-W32, High-Resolution Network, retains high-resolution representations throughout the network, improving spatial precision and robustness to occlusion.

A virtual machine equipped with one NCasT4_v3 GPU powered by an NVIDIA Tesla T4 with 16 GB VRAM was used. The complete pipeline for pose estimation in video footage of group-housed pigs is illustrated in Fig. 5. First, a region of interest (ROI) was defined to isolate the pen within each video frame. Anatomical keypoints were manually annotated in the 1000 selected frames. During annotation, each visible pig was treated as a separate animal instance when individual identity could be determined from the spray-marking pattern and spatial position within the frame. Keypoints were not assigned when the relevant body part was not visible or when the identity of the animal was ambiguous.

For model evaluation, the annotated dataset was randomly divided into 10 repeated training and testing splits using a 95:5 ratio for each split. In each iteration, 950 frames were used for training, and the remaining 50 frames were withheld for testing. This procedure allowed model performance to be evaluated across different subsets of annotated data and reduced dependence on a single train-test partition. In each split, the pose estimation model was trained on the respective training subset and evaluated on the corresponding held-out testing subset. The same data splits were used for all neural network architectures to ensure consistency and comparability across models.

The training process followed a standardized pipeline across models, including data preprocessing, augmentation, and uniform optimization parameters where applicable. These steps included ROI isolation, image resizing and intensity normalization, frame selection and manual annotation, label quality control, data splitting and shuffling, on-the-fly augmentation such as rotations, flips, and brightness or contrast jitter, and batchwise optimization setup. Identical hyperparameters, including batch size, initial learning rate, learning rate schedule, and number of training epochs, were used across networks where supported by the respective DeepLabCut implementation to control for variability. All models were trained using stochastic gradient descent (SGD) with momentum, following the optimization strategy recommended in the official DeepLabCut documentation and the original paper [3]. Learning rates were initialized according to best practices for each architecture and dynamically adjusted during training based on convergence behaviour. The training procedure was run for up to 100,000 iterations, with model convergence confirmed by visual inspection of the training loss curves. All experiments were conducted in Python version 3.9.0 within a custom virtual environment.

2.5. Pose estimation evaluation metrics

Pose estimation performance was evaluated for both experimental stages: the ResNet-50-based comparison of keypoint configurations and the comparison of HRNet-W32, DLCRNet S32, DLCRNet S16, ResNet-50, and ResNet-101. In both cases, model performance was assessed on the training and testing subsets using multiple complementary metrics, including root mean square error (**RMSE**), confidence-filtered RMSE, detection-based errors, mean average precision (**mAP**), and mean average recall (**mAR**). The RMSE was defined as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2]},$$

where N is the total number of keypoints with ground-truth annotations; x_i and y_i are ground-truth coordinates; and $\hat{x}_i$ and $\hat{y}_i$ are predicted coordinates. However, this metric may be misleading if it is based on a

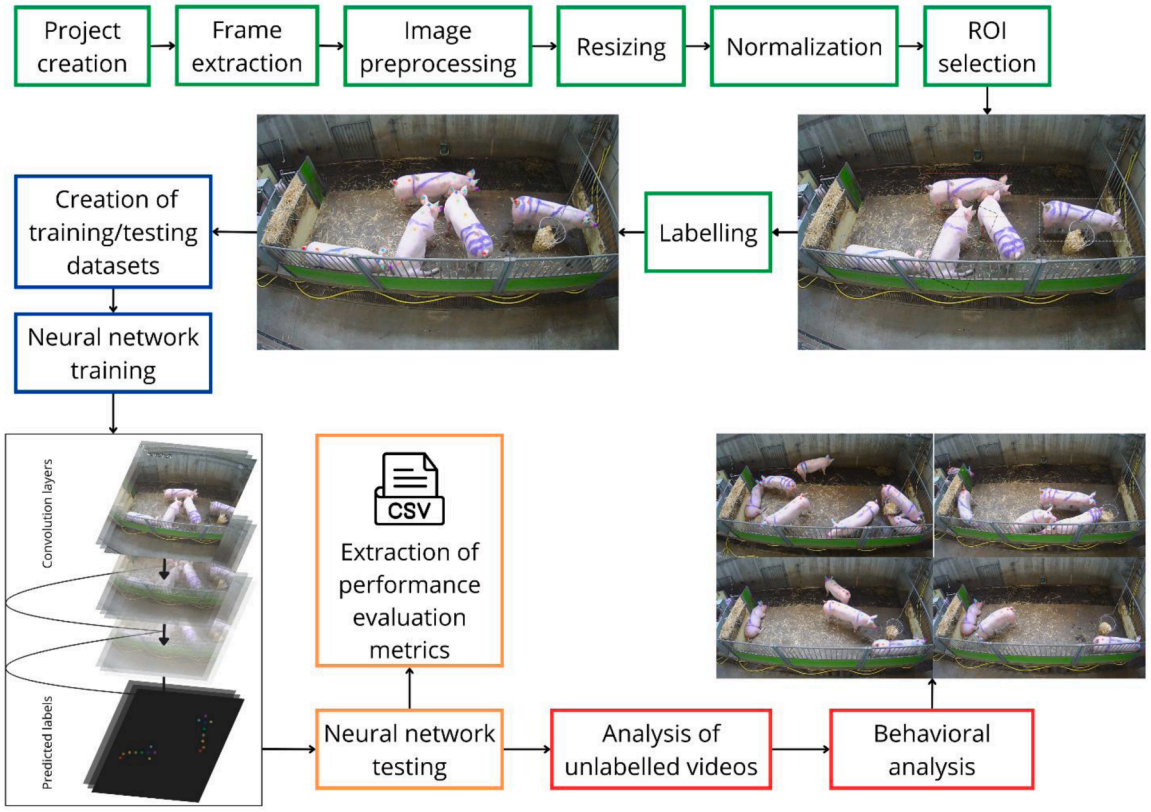

Fig. 5. Pipeline for pose estimation in video footage of group-housed pigs. Green panels represent image preprocessing steps, blue panels correspond to model training procedures, orange panels describe model evaluation, and red panels refer to final video analysis and evaluation for downstream behaviuoral metrics.

large number of keypoint predictions with low confidence.

During inference, each predicted keypoint is assigned a likelihood score (p_i) ranging from 0 to 1, reflecting the model's confidence in the accuracy of that prediction. To ensure reliability, predictions with a confidence score below a predefined threshold (**p_cutoff**) are typically disregarded or masked during post-processing. In our study, we set p_cutoff to 0.6. This threshold was used consistently across the evaluation of keypoint configurations and neural network architectures to enable direct comparison among models. The metric $\mathbf{RMSE_{p\_cutoff}}$ denotes the RMSE calculated only on predicted keypoints that have corresponding ground-truth annotations and meet the confidence criterion ($p_i \geq$ p_cutoff). The $RMSE_{p\_cutoff}$ was calculated as:

$$RMSE_{p\_cutoff} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2]},$$

where M is the number of keypoints with ground-truth annotations and prediction likelihood $p_i \geq$ p_cutoff. This approach offers a more realistic assessment of model accuracy by focusing on predictions made with high confidence. Using $RMSE_{p\_cutoff}$ is particularly valuable in scenarios with frequent keypoint occlusions, where low-confidence predictions are more likely to be erroneous.

Similarly to RMSE and $RMSE_{p\_cutoff}$, we also evaluated $RMSE_{detections}$ and $RMSE_{detections\_p\_cutoff}$. These metrics quantify the RMSE across all keypoints predicted by the model, even when no corresponding ground-truth annotation exists (i.e., including false positive detections). In contrast to standard RMSE metrics, which are computed only for annotated keypoints, $RMSE_{detections}$ and $RMSE_{detections\_p\_cutoff}$ metrics account for all predicted points, providing insight into the model's behaviour in the presence of spurious or extra detections (i.e., detection of points that are not actually present). $RMSE_{detections}$ considers all predictions, while $RMSE_{detections\_p\_cutoff}$ includes only those with $p_i \geq$ p_cutoff.

The *m*AP is the fraction of correctly predicted keypoints among all predicted keypoints. Each average precision is calculated as the area under the precision-recall curve for a specific keypoint, which is built from the calculated precision at multiple thresholds (e.g., different Euclidean distances between predicted and ground-truth keypoints). The *m*AP is calculated as:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} 1(dist_i \leq \varepsilon),$$

where N is the total number of predicted keypoints; $\mathbf{1}(\cdot)$ is the indicator function returning 1 if the condition is true, or 0 otherwise; $dist_i$ is the Euclidean distance between the predicted and ground-truth coordinates for keypoint i; ε is the threshold distance. In this study, $\varepsilon = 5$ px, indicating that a predicted keypoint is considered correct if it falls within 5 px of the ground-truth.

Mean average recall quantifies the proportion of ground-truth keypoints that are correctly predicted by the model. The *m*AR is defined as:

$$mAR = \frac{1}{G}\sum_{i=1}^{G} 1(dist_i \leq \varepsilon \wedge p_i \geq p\_cutoff),$$

where G is the number of ground truth keypoints; $\mathbf{1}(\cdot)$ is the indicator function returning 1 if the condition is true, or 0 otherwise. A prediction is considered correct if the Euclidean distance between the predicted and actual keypoint is within a defined threshold ($\varepsilon = 5$ px), and its confidence score exceeds a specified p_cutoff (0.6).

Based on the aforementioned metrics, the model with the highest overall testing performance was selected and applied to an additional randomly chosen 15-min video that had not been used for training or testing. Keypoint predictions obtained from this video were used to evaluate temporal consistency and reliability of keypoint tracking, as

well as to conduct exploratory spatial behaviour analysis.

2.6. Temporal consistency and reliability of keypoint tracking

To evaluate neural network performance in keypoint tracking, we computed a set of metrics for each keypoint based on data from a randomly chosen 15-min video. This analysis was performed using predictions from the HRNet-W32 model with the BODY5P keypoint configuration, because this combination provided the highest overall testing performance in the preceding comparisons. The 15-min video used for this analysis was obtained from camera 5 and had not been used for training or testing. These metrics were calculated using high-confidence detections from trajectory segments in which the animal instance could be visually identified from the spray-marking pattern and spatial continuity. Segments with missing detections, low-confidence predictions, severe overlap, or uncertain identity assignment were treated as discontinuous and were not used for identity-dependent temporal calculations. Therefore, velocity and jerk should be interpreted as measures of temporal stability within reliable trajectory segments rather than as indicators of uninterrupted long-term individual tracking. The analysis was restricted to pig body parts with more than 10,000 valid detections (i.e., detections with $p_i \geq$ p_cutoff) to ensure statistical robustness. To evaluate the temporal consistency and reliability of keypoint tracking, three dynamic tracking metrics were analyzed: keypoint detection failure rate (dropout), keypoint velocity, and jerk. Dropout quantifies the proportion of keypoint predictions that fall below the specified confidence threshold. Dropout was calculated separately for each anatomical keypoint as the proportion of predictions with likelihood below the confidence threshold p_{cutoff} over the total number of predictions for that keypoint across the dataset. This per-keypoint dropout rate quantifies the frequency with which individual body parts are missed or detected with low confidence, providing insight into model robustness against occlusions or labelling difficulties. A high dropout rate suggests the keypoint is often occluded, poorly labeled, or challenging to detect. This metric is critical when evaluating model performance, especially in settings where certain body parts are frequently obscured or ambiguous.

Keypoint velocity represents the change in position of a keypoint between consecutive frames (inter-frame displacement, px/frame). It reflects the typical movement magnitude of each keypoint and provides insight into expected motion dynamics. High average velocity suggests that the keypoint is associated with a body part that moves frequently or rapidly. Consistent velocity (with low standard deviation) implies smooth, predictable motion and reliable tracking over time.

Jerk, defined as the change in velocity between consecutive frames (first derivative of velocity expressed as px/frame^2, i.e., how quickly the speed or direction of motion is changing), was used to assess the smoothness of keypoint trajectories over time. It indicates how abruptly the movement of a keypoint is changing. Together, these metrics offer a comprehensive view of the spatiotemporal behaviour of each tracked keypoint, highlighting strengths in stable keypoint localization and revealing weaknesses in regions prone to tracking interruptions or noise.

2.7. Spatial behaviour analysis

Keypoint predictions from the 15-min video used to evaluate temporal consistency were also used to explore the spatial dynamics of fattening pigs in our setup. The same HRNet-W32 model and BODY5P configuration were used for this analysis. This analysis was exploratory and was intended to assess whether pose-estimation outputs could provide interpretable spatial indicators of animal activity and potential social interactions. The analysis was restricted to pig body parts with more than 10,000 valid detections, defined as detections with prediction likelihood pi $\geq$ 0.9. A higher likelihood threshold was used for this exploratory spatial analysis to reduce the risk of false-positive proximity events.

A two-dimensional occupancy heatmap was generated based on predicted body landmarks to visualize patterns of space use and movement within the pen. The heatmap was calculated from high confidence keypoint detections and used to identify areas with higher or lower spatial activity across the analysed video segment.

In addition, we analysed tail-snout proximity events as candidate indicators of tail-directed social interaction. For each frame, pairwise Euclidean distances were calculated between high-confidence snout detections and high-confidence tail-base detections belonging to different animal instances. Snout-tail pairs from the same animal were excluded. A proximity event was defined when the snout of one pig was located within a predefined distance of the tail base of another pig. Two pixel-based thresholds were evaluated: 50 px as a broader exploratory proximity criterion and 20 px as a stricter criterion for close-contact events. Frames with missing, low-confidence, or identity-ambiguous snout or tail detections were excluded from this analysis.

Because these thresholds were defined in image pixels and were not calibrated to physical distance, the resulting proximity events were interpreted as exploratory spatial indicators rather than confirmed tail-biting events. This analysis was therefore used to demonstrate the potential of pose-estimation outputs for identifying spatial zones and candidate interaction events that may warrant further behaviour-specific validation.

Table 1
Performance of five keypoint configurations for automated pose estimation in pigs. The configurations included a 7-keypoint model (STD) and four variants with additional anatomical landmarks focused on the ears (EAR3P), body (BODY5P), eyes (EYES2P), and tail (TAIL2P).

Performance metrics[1]	STD	EAR3P	BODY5P	EYES2P	TAIL2P
Training					
RMSE (px)	1.89 ± 0.06	2.24 ± 0.08	2.36 ± 0.09	2.16 ± 0.07	2.39 ± 0.09
$RMSE_{p\_cutoff}$ (px)	1.88 ± 0.06	2.13 ± 0.07	2.35 ± 0.08	2.06 ± 0.06	2.29 ± 0.07
*m*AP (%)	96.71 ± 0.41	98.05 ± 0.35	98.62 ± 0.36	97.84 ± 0.27	96.18 ± 0.44
*m*AR (%)	99.82 ± 0.05	99.84 ± 0.05	99.59 ± 0.03	99.25 ± 0.06	99.77 ± 0.05
$RMSE_{detections}$ (px)	1.85 ± 0.07	2.23 ± 0.08	2.37 ± 0.08	2.45 ± 0.08	2.11 ± 0.07
$RMSE_{detections\_p\_cutoff}$ (px)	1.85 ± 0.06	2.14 ± 0.07	2.37 ± 0.08	2.04 ± 0.06	1.98 ± 0.06
Testing					
RMSE (px)	13.21 ± 0.74	52.30 ± 3.15	7.43 ± 0.43	13.88 ± 0.67	17.05 ± 0.82
$RMSE_{p\_cutoff}$ (px)	8.61 ± 0.51	6.14 ± 0.38	6.12 ± 0.35	10.94 ± 0.58	5.74 ± 0.34
*m*AP (%)	68.40 ± 1.78	51.88 ± 2.24	75.12 ± 1.42	72.10 ± 1.58	61.45 ± 2.05
*m*AR (%)	78.20 ± 1.56	59.50 ± 2.10	90.15 ± 1.18	77.30 ± 1.62	67.80 ± 1.98
$RMSE_{detections}$ (px)	12.04 ± 0.68	44.52 ± 2.73	7.05 ± 0.41	18.37 ± 0.93	27.16 ± 1.28
$RMSE_{detections\_p\_cutoff}$ (px)	8.74 ± 0.46	26.85 ± 1.62	5.74 ± 0.33	12.43 ± 0.68	17.82 ± 0.88

[1] RMSE: root mean square error computed over all keypoints with available ground-truth annotations; $RMSE_{p\_cutoff}$: RMSE computed only for keypoints with a prediction likelihood $\geq$ p_cutoff; *m*AP: mean average precision, i.e. the percentage of predicted keypoints within a 5-px radius of ground truth, regardless of ground-truth presence; *m*AR: mean average recall: the percentage of ground-truth keypoints correctly predicted within a 5-px radius and above the likelihood threshold (p_cutoff); $RMSE_{detections}$: RMSE computed over all predicted keypoints, including false positives, regardless of prediction likelihood; $RMSE_{detections\_p\_cutoff}$: RMSE computed over all predicted keypoints with a likelihood $\geq$ p_cutoff, including those not matched to ground-truth annotations.

3. Results and discussion

3.1. Pose estimation performance across keypoint configurations

Table 1 summarizes the performance of five keypoint configurations tested to identify the most effective skeletal model for automated pose estimation in pigs. The configurations included a 7-keypoint model (standard model, STD) and four variants with additional anatomical landmarks focused on the ears (EAR3P), body (BODY5P), eyes (EYES2P), and tail (TAIL2P). All anatomical landmarks were manually annotated on an expanded set of 500 images. For each configuration, 400 images were used for training, while the remaining 100 images, also manually annotated but withheld from training, served as a held-out test set to evaluate model generalization. The same train-test split was applied across all configurations to ensure that differences in performance reflected the skeletal model design rather than differences in image content.

Among the configurations, BODY5P consistently demonstrated superior generalization, achieving the highest test *m*AP (75.12 ± 1.42%) and the lowest overall test RMSE (7.43 ± 0.43 px), indicating robust pose estimation performance. Additionally, BODY5P yielded the lowest $RMSE_{detections}$ (7.05 ± 0.41 px) and the lowest $RMSE_{detections\_p\_cutoff}$ (5.74 ± 0.33 px), suggesting minimal false positives and strong spatial accuracy when predictions were above the confidence threshold (p_cutoff = 0.6). BODY5P also achieved the highest test *m*AR (90.15 ± 1.18%), indicating that a larger proportion of ground-truth keypoints was correctly recovered in the held-out test set. These results suggest that multiple anatomically distributed points along the dorsal body axis provide a stable and occlusion-resistant framework for tracking pigs in group-housing conditions.

In contrast, TAIL2P and EYES2P, while achieving high accuracy during training (training mAP 96.18 ± 0.44% and 97.84 ± 0.27%, respectively), showed a decline in test performance, likely due to frequent occlusions or loss of visibility of these small, mobile features during natural behaviour. EYES2P still showed relatively high test *m*AP (72.10 ± 1.58%), but its higher test RMSE (13.88 ± 0.67 px) and $RMSE_{detections}$ (18.37 ± 0.93 px) indicate less stable localization when all predictions and detections were considered. TAIL2P showed lower overall test *m*AP (61.45 ± 2.05%) and *m*AR (67.80 ± 1.98%), suggesting less consistent detection across the full test set. However, TAIL2P achieved the lowest $RMSE_{p\_cutoff}$ (5.74 ± 0.34 px), indicating that when high-confidence tail-related predictions were available, their localization could be spatially accurate.

Although the TAIL2P and EYES2P configurations exhibited lower overall pose estimation accuracy and higher variability due to frequent occlusions and rapid movements of these regions, they capture behaviourally relevant features not represented in other skeletal models. This suggests their utility in specialized tasks such as detecting tail-biting or ear-nibbling behaviours, which warrant dedicated investigation despite current performance trade-offs. In this sense, lower overall localization performance does not necessarily imply lower biological relevance. Rather, it highlights the challenge of reliably detecting anatomically small or highly mobile landmarks that may be crucial for specific behavioural endpoints.

The EAR3P configuration performed the worst overall, with a test RMSE of 52.30 ± 3.15 px and *m*AP of 51.88 ± 2.24%, reflecting the high susceptibility of ear keypoints to occlusion, particularly during close interactions or overlapping body postures. EAR3P also showed the lowest test *m*AR (59.50 ± 2.10%) and the highest $RMSE_{detections\_p\_cutoff}$ (26.85 ± 1.62 px), indicating that additional ear landmarks introduced substantial uncertainty under group-housing conditions. The difference between training and testing $RMSE_{detections\_p\_cutoff}$ values further indicates model overfitting and sensitivity to noisy or ambiguous keypoints. While all configurations had relatively low RMSE values during training, BODY5P showed the most stable metrics across training and testing, reinforcing its robustness and anatomical reliability. Conversely, configurations such as TAIL2P, EYES2P, and especially EAR3P exhibited greater variability and intermittent keypoint loss, likely due to reduced visibility, rapid movement, and occlusion of these anatomical regions.

These results emphasize the pivotal role of skeletal model design in optimizing pose estimation for fattening pigs. The superior performance of BODY5P suggests that multiple, anatomically distributed points along the pig's back provide a stable and occlusion-resistant framework for tracking in group-housing conditions. However, to capture detailed biologically relevant behaviours like tail biting or ear nibbling, model configurations need to include keypoints that, while challenging to detect reliably, offer high behavioural resolution. Therefore, the configuration that achieves the best overall localization accuracy is not necessarily the configuration that is most suitable for every downstream behavioural task. BODY5P should be interpreted as the most robust configuration for general pose estimation in this dataset, whereas behaviour-specific applications such as tail-biting detection require particular attention to the reliability of snout and tail-related landmarks.

In comparison, commonly used object detection frameworks such as YOLO are highly efficient for locating and segmenting whole animals but are limited in providing fine-grained, anatomically detailed information necessary for pose-based behaviour analysis. The skeletal-based approach applied in this study therefore complements and extends YOLO-style detection by enabling the quantification of individual body part dynamics. Future work should focus on improving detection robustness for behaviourally relevant but visually challenging keypoints through targeted annotation, more diverse training data, confidence threshold optimization, and post-processing approaches that balance spatial precision with biological relevance. For tail-directed behaviours specifically, future analyses should evaluate keypoint configurations using task-specific criteria, including snout and tail localization error, dropout rate of tail-related landmarks, and agreement with manually annotated tail-directed interactions.

These findings are consistent with previous studies that emphasized the importance of selecting well-defined and consistently visible keypoints. For example, recent work by Liu et al. [9] achieved high tracking accuracy in piglets using a deep learning-based skeleton model and highlighted the value of spatially distinct, behaviourally relevant keypoints for pose tracking. Similarly, recent studies on sow locomotion have demonstrated that skeletal models incorporating 6 to 10 anatomical landmarks achieve optimal performance for locomotion analysis and behaviour monitoring [6]. These findings align closely with our standard (STD) and BODY5P configurations, which balance anatomical coverage and robustness required for precise pig pose estimation. Finally, our results also highlight the limitations of using keypoints prone to frequent occlusion or high motion blur, such as snouts, eyes, ears, and tail-related landmarks. The significant performance drop in these configurations underlines the challenge of maintaining prediction reliability in complex, multi-animal environments.

3.2. Neural network performance

The performance metrics for pose estimation of the five neural networks compared in this study (HRNet-W32, DLCRNet S32, ResNet-50, ResNet-101, and DLCRNet S16) revealed differences in accuracy and generalization ability (Table 2). This comparison was performed using the selected BODY5P keypoint configuration and identical data splits across architectures, allowing the effect of neural network backbone to be assessed independently of differences in landmark definition or image partitioning. Training performance showed that HRNet-W32 achieved superior accuracy, with the lowest RMSE (1.96 ± 0.11 px) and near-perfect localization metrics (*m*AP = 99.80%, *m*AR = 99.96%). These results highlight the model's strong capability to learn accurate spatial details, likely due to its architecture, which maintains high-resolution feature maps throughout the network. In contrast, ResNet-50 and DLCRNet variants (S16 and S32) exhibited higher training RMSE values (> 6 px), along with reduced precision (*m*AP between

Table 2
Performance metrics obtained across different neural networks for pig pose estimation (HRNet-W32, DLCRNet S32, ResNet-50, ResNet-101, and DLCRNet S16).

Performance metrics[1]	HRNet-W32	DLCRNet S32	ResNet-50	ResNet-101	DLCRNet S16
Training					
RMSE (px)	1.96 ±0.11	5.99 ±0.67	6.36 ±1.13	5.06 ±0.56	6.13 ±0.41
$RMSE_{p\_cutoff}$ (px)	1.96 ±0.11	4.80 ±0.28	4.23 ±0.26	4.03 ±0.41	4.54 ±0.29
*m*AP (%)	99.80 ±0.08	92.82 ±0.94	91.73 ±1.54	93.95 ±0.79	91.94 ±0.67
*m*AR (%)	99.96 ±0.03	93.98 ±0.95	93.03 ±1.27	95.09 ±0.60	93.26 ±0.65
$RMSE_{detections}$ (px)	1.95 ±0.11	5.80 ±0.28	5.07 ±0.36	4.64 ±0.46	5.24 ±0.35
$RMSE_{detections\_p\_cutoff}$ (px)	1.95 ±0.11	4.66 ±0.21	3.63± 0.19	3.69 ±0.18	3.93 ±0.20
Testing					
RMSE (px)	7.93 ±0.36	8.91 ±1.24	10.53 ±1.43	8.67 ±0.76	8.76 ±1.30
$RMSE_{p\_cutoff}$ (px)	6.68 ±0.20	7.13 ±0.40	7.54 ±1.21	6.84 ±0.21	7.52 ±1.42
*m*AP (%)	92.52 ±0.96	89.93 ±1.70	88.15 ±1.38	88.77 ±1.98	90.59 ±0.96
*m*AR (%)	93.59 ±0.70	90.95 ±1.57	89.50 ±1.37	90.18 ±1.80	91.55 ±1.01
$RMSE_{detections}$ (px)	8.04 ±0.32	8.15 ±0.70	7.27 ±0.46	7.91 ±0.62	7.68 ±0.26
$RMSE_{detections\_p\_cutoff}$ (px)	6.85 ±0.20	7.02 ±0.27	6.70 ±0.16	6.75 ±0.30	7.00 ±0.24

[1] RMSE: root mean square error computed over all keypoints with available ground-truth annotations; $RMSE_{p\_cutoff}$: RMSE computed only for keypoints with a prediction likelihood ≥ p_cutoff; *m*AP: mean average precision, i.e. the percentage of predicted keypoints within a 5-px radius of ground truth, regardless of ground-truth presence; *m*AR: mean average recall: the percentage of ground-truth keypoints correctly predicted within a 5-px radius and above the likelihood threshold (p_cutoff); $RMSE_{detections}$: RMSE computed over all predicted keypoints, including false positives, regardless of prediction likelihood; $RMSE_{detections\_p\_cutoff}$: RMSE computed over all predicted keypoints with a likelihood ≥ p_cutoff, including those not matched to ground-truth annotations.

91.7% and 92.8%). Although ResNet-101 showed moderate training RMSE (5.06 ± 0.56 px), it outperformed ResNet-50 and the DLCRNet models in *m*AP (93.95%) and *m*AR (95.09%), indicating more accurate learning of keypoint positions detected during training.

Testing performance further emphasized the advantage of HRNet-W32, which maintained the best overall RMSE (7.93 ± 0.36 px) and RMSEp_cutoff (6.68 ± 0.20 px), with the highest generalization accuracy (*m*AP = 92.52%, *m*AR = 93.59%). These findings support its robustness in handling occlusions and spatial variability in unseen data, consistent with previous work showing the advantages of high-resolution representations in multi-animal and occluded contexts [10]. The DLCRNet variants, which were designed for computational efficiency, demonstrated slightly lower accuracy overall but still performed reasonably well in test settings (e.g., DLCRNet S32 test RMSE = 8.91 ± 1.24 px, *m*AP = 89.93%). Although ResNet-50 achieved the lowest RMSEdetections_p_cutoff during testing (6.70 ± 0.16 px), the differences across all models were relatively small when low-confidence predictions were filtered out. This suggests that, under high-confidence prediction conditions, several architectures could produce spatially accurate detections, even if their overall generalization performance differed.

Across the evaluated neural network architectures, all models demonstrated generally strong capabilities for keypoint detection in group-housed pigs, with only modest differences in predictive robustness. In particular, when model predictions were filtered to include only high-confidence detections, the variation in performance across networks became less pronounced. However, it is noteworthy that the SE associated with performance metrics varied across architectures. For example, DLCRNet S32 and ResNet-50 showed higher variability in training RMSE (e.g., 5.99 ± 0.67 px and 6.36 ± 1.13 px, respectively) and test RMSE (e.g., 8.91 ± 1.24 px and 10.53 ± 1.43 px), compared with HRNet-W32 (7.93 ± 0.36 px). These differences suggest that certain models may be more sensitive to data partitioning or training conditions, reflecting less consistent performance across different data splits.

However, selecting the most suitable pose estimation model should not rely solely on traditional accuracy metrics. The practical value of a deep learning system for livestock monitoring is shaped by multiple other factors, including robustness to occlusion, generalizability to new farm conditions, computational efficiency, inference speed, and the biological relevance of predicted keypoints. Reviews of precision livestock farming technologies have emphasized that model selection in real-world animal monitoring should account for the consistency of keypoint detection under variable lighting, occlusions due to group behaviour, and shifting pen configurations, which are challenges not typically captured by accuracy alone [11,12].

Architectures originally designed for high-resolution spatial tasks, such as HRNet, offered precise keypoint localization, which is particularly valuable in contexts where occlusion and animal overlap are common. Nonetheless, deeper networks such as ResNet-101 and optimized variants like DLCRNet models also performed competitively, particularly when detection confidence was considered. Previous studies have shown that deeper networks can improve performance in complex visual tasks [13]. In animal behaviour research, ResNet-based architectures have been widely adopted for their balance of accuracy and efficiency [3,14]. However, direct performance comparison across studies is not straightforward, as results depend heavily on dataset composition, image resolution, species, camera angle, annotation strategy, and evaluation protocol. Our findings are consistent with this general trend: increasing network depth and feature complexity can improve learning stability, but well-configured lighter architectures such as DLCRNet can achieve comparable accuracy under practical farm conditions. This suggests that the optimal choice of backbone depends not only on absolute performance metrics but also on computational constraints and environmental variability.

Training loss trajectories (Fig. 6) indicated that all models converged consistently over 100 epochs, with HRNet-W32 achieving the lowest loss with smooth and stable optimization, indicating its superior capacity to model complex pose patterns. ResNet architectures also demonstrated reliable convergence, although with slightly higher losses and greater variance during early epochs, particularly for the deeper ResNet-101, reflecting the challenges deeper networks face with optimization and hyperparameter tuning. DLCRNet models converged fastest but plateaued at higher final losses, consistent with their comparatively lower accuracy metrics like RMSE and *m*AP. This underscores the familiar trade-off between computational efficiency and predictive precision.

Model training times varied substantially across neural networks, with HRNet-W32 requiring 4.3 to 7.5 times longer than DLCRNet to complete. ResNet-50 and ResNet-101 occupied an intermediate position, taking 2 to 3 times longer than DLCRNet to run (data not shown). These differences reflect the inherent computational complexity of each architecture: HRNet, for instance, maintains high-resolution representations throughout its layers, which greatly benefits spatial precision but increases training cost [10]. Hence, although HRNet-W32 achieved the highest accuracy and the most consistent performance across both training and test sets, it did so at a considerable computational cost. In contrast, DLCRNet models, while yielding slightly lower accuracy and higher error variability, trained substantially faster and still delivered competitive performance in test conditions, making them attractive candidates for rapid prototyping, particularly in resource-constrained settings or applications requiring frequent retraining. This aligns with previous work in animal pose estimation, which has highlighted the benefits of lightweight architectures for enabling scalable, real-time,

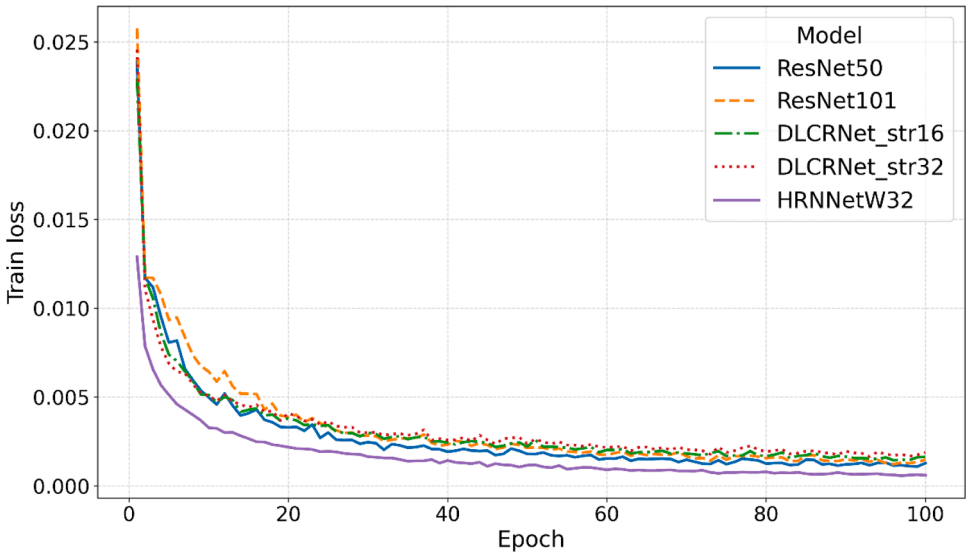

Fig. 6. Loss of the five neural networks across epochs during model training.

and hardware-efficient deployments without incurring substantial losses in predictive utility [15,16]. ResNet-50 offered a favorable trade-off between training time and generalization performance, with consistently low prediction variability, positioning it as a practical baseline for moderate-resource settings.

These results highlight that selecting a pose estimation architecture requires balancing accuracy, prediction stability, and computational demands. For example, in high-throughput farm monitoring systems or embedded deployments, where inference speed is prioritized over marginal gains in accuracy, DLCRNet variants offer a compelling trade-off, enabling rapid retraining and efficient deployment without significant sacrifices in model robustness. Their performance is particularly suitable also for individual-animal tracking tasks, where spatial complexity and instance disambiguation are minimal. Their lightweight design makes them ideal for deployment in field studies or embedded systems with limited hardware capabilities. Conversely, in research contexts or precision phenotyping scenarios where fine-grained accuracy and consistency are paramount, HRNet-W32 remains the preferred option, despite its heavier computational demands. This context-dependent approach to model choice reflects growing consensus in the field [3,14,15,17], ensuring that pose estimation tools remain both practical and scalable for diverse animal behaviour research settings.

Given the superior accuracy of HRNet-W32, this architecture was selected for pose estimation in a randomly chosen 15-min video to test temporal consistency and reliability of keypoint tracking and to perform spatial behaviour analysis.

3.3. *Temporal consistency and reliability of keypoint tracking*

Metrics of temporal consistency and reliability of keypoint tracking are reported, for each keypoint, in Table 3. These metrics were interpreted using high-confidence trajectory segments in which the animal instance could be identified from the spray-marking pattern and spatial continuity. Segments with missing detections, severe overlap, or uncertain identity assignment were treated as discontinuous. Therefore, velocity and jerk should be interpreted as measures of temporal stability within reliable trajectory segments rather than as evidence of uninterrupted long-term individual tracking across all occlusion events. Dropout rates for the head and most dorsal body landmarks ranged from approximately 18% to 29%, indicating relatively stable detection for several core body regions. Higher dropout was observed for Body3 (31.6%), the right ear (32.9%), and the tail (31.8%), whereas the snout exhibited the highest dropout rate of 46.0%. This reflected difficulties in reliable localization due to occlusions, head orientation changes, and limited frontal view representation in the training data. Nevertheless, the snout remains a critical anatomical landmark for behavioural analysis because it is essential for capturing facial behaviour and tail-directed interactions, despite the increased detection challenges. While direct studies focusing on snout detection difficulties in pigs are limited, several recent works highlight challenges in accurate facial landmark localization in group-housed pigs due to occlusion and pose variations (Jayachitra [4,1]). These findings suggest the need for improved annotation strategies and model architectures tailored for pig facial keypoints. Mean velocities varied from ~17 px/frame for the tail and snout to ~26 px/frame for Body2 and Body4, aligning with expected locomotor dynamics at the given video resolution and frame rate. The standard deviation of velocity was lowest for the snout and tail (7–8 px), suggesting relatively consistent motion when these points were visible, whereas the elevated variability for Body2 and Body4 (10 px) likely reflects tracking instabilities, such as abrupt keypoint displacements, misdetections, or jitter, phenomena also observed in studies using DeepLabCut and LEAP for animal pose tracking, particularly during turning or fast movement sequences [5,16]. Jerk values were on the order of 10^{-4} px/frame^2 with minimal standard deviation, indicating smooth motion trajectories in most instances, despite occasional outliers. These results collectively highlight both the strengths and limitations of the current pose estimation model. Low dropout rates and stable

Table 3
Keypoint tracking performance metrics[1].

Keypoint	Dropout, %	Velocity, px/frame	Jerk, px/frame^2
Snout	46.0	18.83	0.000054
Head	18.5	24.25	0.000074
Left ear	28.8	23.57	0.000051
Right ear	32.9	23.33	0.000085
Body1	18.3	24.34	0.000053
Body2	25.9	25.95	0.000048
Body3	31.6	24.94	0.000033
Body4	27.4	25.33	0.000083
Body5	29.0	21.21	0.000028
Tail	31.8	16.97	0.000054

[1] Dropout: proportion of frames in which a given keypoint is not detected because its predicted confidence (likelihood) falls below a predefined threshold (p_cutoff); Velocity: mean frame-to-frame displacement of the keypoint across time, expressed in pixels per frame; Jerk: average magnitude of change in velocity between consecutive frames (i.e., first derivative of velocity).

motion patterns for keypoints like Body1, head, and ears demonstrate the model's strong ability to reliably track anatomically rigid and visually distinct regions. Conversely, the high dropout rate for the snout emphasizes the need for enhanced training diversity, including a greater variety of frontal and oblique facial poses (e.g., by increasing the number of cameras or placing cameras at different angles), and possibly lowering the detection confidence threshold or employing targeted data augmentation strategies to improve robustness.

The higher velocity variability observed for Body2 and Body4 further suggests the utility of post-processing techniques, such as Savitzky–Golay smoothing or Kalman filtering [17,18], to mitigate abrupt coordinate fluctuations without compromising the fidelity of actual motion. Although the tail demonstrated moderate dropout, its low jerk and consistent velocity patterns indicate that, when detected, tail keypoints are tracked reliably. Strategic annotation of occluded tail frames may further enhance detection frequency, especially given the behavioural relevance of tail motion in swine studies [19]. In sum, while the model performs effectively for core body landmarks, targeted improvements are warranted for enhanced tracking of facial and tail dynamics, which is critical for downstream behavioural analyses such as the analysis of candidate tail-directed interactions.

3.4. Spatial behaviour analysis

Understanding the spatial distribution and locomotor patterns of animals is fundamental in behavioural ethology, particularly within intensive farming systems where welfare issues such as tail biting may emerge due to environmental constraints, social stress, or lack of enrichment. Longitudinal analysis of position data enables researchers to identify preferential zones, recurrent locomotor patterns, and inactivity, which serve as proxies for behavioural and welfare monitoring [20,21].

Analysis of spatial behaviour based on 2D heatmaps (Fig. 7) revealed pronounced asymmetries in pen occupancy. The left side of the region of interest exhibited the highest concentration of occupancy, suggesting repeated visits or prolonged occupancy in that area. This corresponds to the area in proximity of the fixed straw baskets. In contrast, the central and right regions of the region of interest showed more diffuse and transient activity patterns, which may reflect exploratory behaviour or avoidance of dominant individuals. These findings align with previous observations of environmental and social drivers of asymmetrical space use in pigs [22], highlighting the role of both physical and social structures in shaping group movement and area preference. As reported in Valros et al. [23], clustering of activity in specific zones may indicate the presence of environmental stressors, resource competition, or elevated social tension. Additionally, in line with our results, occupancy hotspots often align with enrichment zones, such as manipulable objects or straw, feeding areas, or socially preferred locations where interactions occur more frequently. Such clustering may reflect both positive social interactions and potential welfare concerns, including tail biting episodes or aggressive encounters [24].

Beyond descriptive insight, pose-based spatial tracking provides a non-invasive, scalable method for monitoring group-housed animals. In the context of video-based animal monitoring, heatmaps offer an interpretable and compact representation of spatial behaviour over time, serving as a bridge between raw pose estimation data and higher-level behavioural analysis [15,17]. These visualizations can be used not only to quantify space use but also as input features for downstream applications, such as unsupervised clustering, behaviour classification, or anomaly detection via convolutional neural networks [25]. For instance, prior studies have integrated spatial occupancy metrics into deep learning models to detect stereotypic behaviour, resting states, and episodes of aggression across various species [26]. In this context, the heatmaps generated in our study serve a dual role: first, as intuitive visualizations of movement patterns and spatial preferences; and second, as foundational data representations for the development of predictive models aimed at identifying welfare-relevant behaviours, including resting, aggression, and stereotypies [26].

Recent advances in pig behaviour analysis have demonstrated that spatial and temporal representations are highly effective for quantifying complex interactions and behavioural rhythms. For example, Gan et al. [27] developed a key point-based spatial and temporal feature extraction approach to automatically detect and analyse social behaviours among preweaning piglets, showing that spatial distributions of keypoints can capture affiliative and aggressive interactions with high accuracy. Similarly, Huang et al. [28] proposed a behaviour classification and rhythm analysis framework based on wearable ear-tag sensors, using time-amplitude and density plots to reveal feeding, movement, and resting patterns in fattening pigs. These studies collectively support the relevance of heatmap-based representations as interpretable tools for visualizing behavioural organization and as informative inputs for predictive modelling of welfare-related states in group-housed pigs.

To illustrate the potential of neural network-based pose estimation for analysing spatial aspects of social interactions, particularly those relevant for the future detection of tail-directed behaviours, we examined tail-snout proximity events. These events were defined as video frames in which the snout of one pig was located within a specified distance from the tail base of another pig. In accordance with the methodological criteria described above, proximity events were calculated only from high-confidence detections and only between different

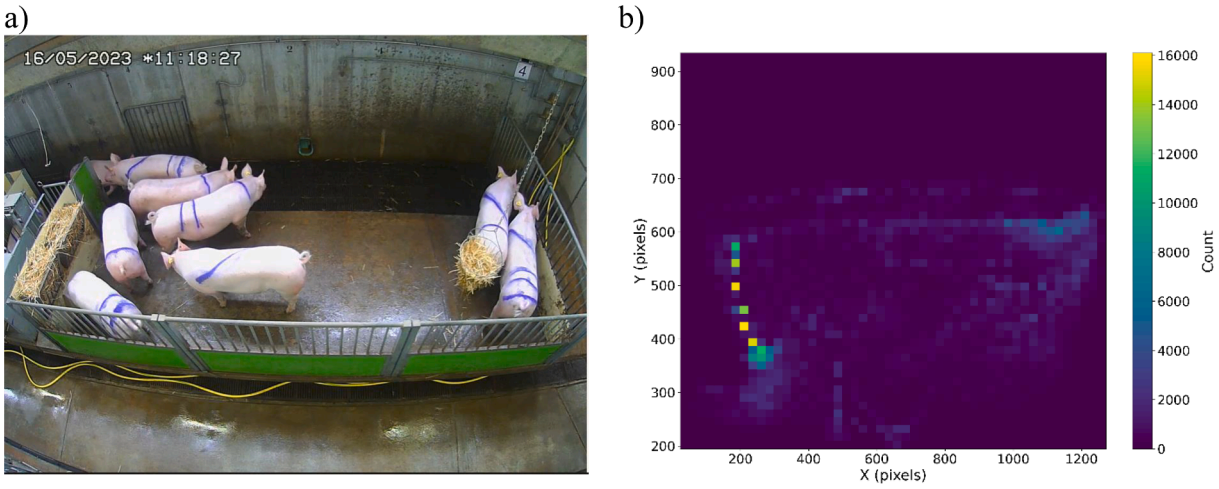

Fig. 7. Frame of the pen (a) and corresponding 2D occupancy map (b) for all pigs in the zone of the pen captured by the video recordings.

animal instances. Snout-tail pairs belonging to the same individual were excluded, and frames with missing, low-confidence, or identity-ambiguous detections were not used for this analysis.

Fig. 8a visualizes such interactions using a 50-px threshold, while Fig. 8b applies a stricter 20-px threshold to isolate closer proximity events. The 50-px threshold should be interpreted as a broader exploratory criterion for identifying possible spatial attention toward the tail region, whereas the 20-px threshold represents a stricter criterion for close-contact candidate events. In both cases, we observed spatial clustering of tail-snout proximity events, particularly near pen boundaries and previously identified hotspots in the occupancy heatmaps (Fig. 7). This spatial overlap suggests that candidate tail-directed proximity events tend to co-occur in zones of elevated occupancy or repeated spatial use, such as feeding areas, enrichment areas, or near structural barriers.

Although these events do not constitute definitive evidence of tail biting, they may serve as candidate indicators of social attention toward the tail region. This distinction is important because spatial closeness alone cannot determine motivation, contact type, or behavioural outcome. A snout located near another animal's tail may reflect exploratory investigation, incidental co-location, feeding-related crowding, avoidance behaviour, or genuine tail-directed interaction. Therefore, tail-snout proximity should be interpreted as an exploratory spatial indicator rather than as a validated behavioural classification of tail biting.

Despite its utility, this proximity-based method has inherent limitations. First, spatial closeness alone does not imply intentional interaction; shared occupancy may result from pen design rather than social preference. Second, the use of fixed pixel thresholds does not account for individual body size variation, body orientation, or camera perspective distortions. Additionally, pose estimation errors in group-housed settings, especially during occlusion or close contact, can lead to false positives, where proximity is erroneously inferred. To mitigate these challenges, recent approaches have advocated for integrating temporal pose trajectories and behaviour classification algorithms to better differentiate between incidental proximity and deliberate social actions [2,29–31]. Future work should therefore combine proximity-based indicators with manually annotated behavioural events, temporal persistence criteria, and possibly adaptive distance thresholds normalized by body size or calibrated physical distance.

Nonetheless, proximity mapping remains a valuable first step in identifying behaviourally salient spatial zones and candidate interaction events that may warrant further manual validation. As part of a broader pose-based monitoring pipeline, this approach may help prioritize video segments for review, identify spatial hotspots of social attention, and provide input features for future supervised models targeting welfare risks such as tail biting in intensive pig production.

3.5. Limitations of the study

While the present study provides a detailed and methodologically rigorous evaluation of different keypoint configurations and neural network architectures for pose estimation in pigs, several important limitations must be acknowledged. Addressing these constraints is crucial for translating pose estimation models into fully deployable and behaviour-aware systems in practical farm settings.

Firstly, although the annotated dataset was expanded for the keypoint-configuration comparison and a separate 1000-frame dataset with repeated training and testing splits was used for the neural network architecture comparison, the overall diversity of the annotated material remains limited relative to the variability expected in commercial pig production. The keypoint-configuration comparison was based on 500 annotated images, with 400 images used for training and 100 withheld for testing, while the architecture comparison used 1000 annotated frames and a repeated train-test evaluation strategy. These additions strengthen the reliability of the comparisons compared with a very small held-out test set. Nevertheless, both experimental stages were conducted within the same farm environment, using the same group of animals and the same camera perspective. Therefore, the range of postures, lighting conditions, group compositions, and body orientations captured in the annotated data may still remain constrained. This limitation is particularly relevant in group-housed environments, where frequent occlusions, overlaps, and complex interactions are prevalent. As shown in recent work by Jeon et al. [32], pose estimation models trained on small or homogeneous datasets are vulnerable to overfitting and may exhibit reduced robustness under dynamic conditions. Their findings emphasized the value of larger, behaviourally diverse datasets, ideally annotated under multiple conditions and across different animal cohorts, to enhance generalization and reduce prediction drift in natural environments [32].

Secondly, the current study relied on a single overhead camera for training and evaluation. While this setup is cost-effective and commonly used in livestock applications, it introduces a limitation in spatial perception, especially for tracking fine movements and occluded body parts. Overhead-only views are often insufficient to capture certain behaviours, such as facial interactions or low-angle tail postures, particularly during close group contact. This limitation is directly relevant to the observed lower reliability of small or mobile landmarks such as the

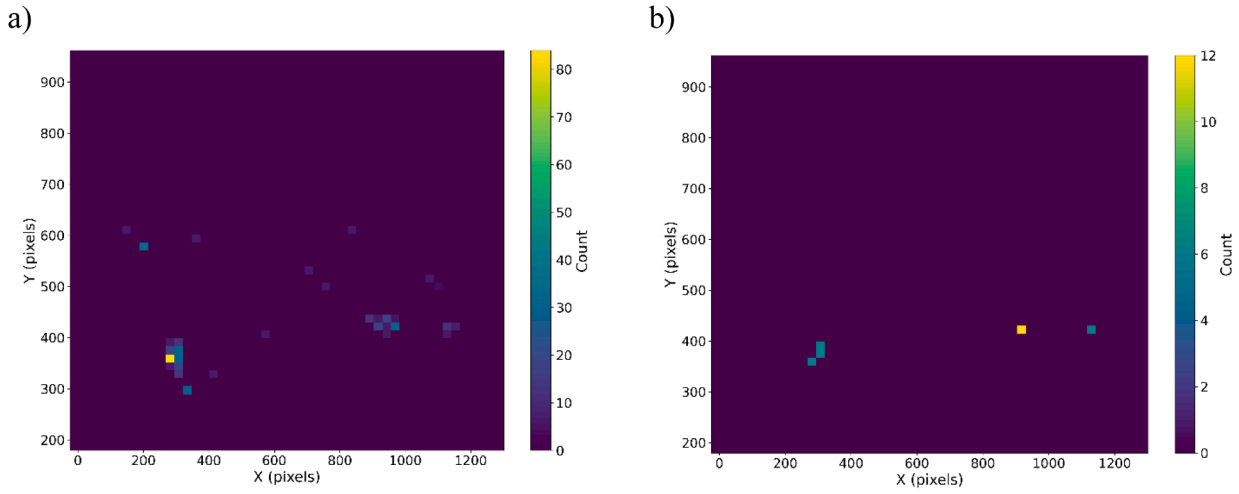

Fig. 8. Heatmaps of tail-snout proximity events, defined as video frames in which the snout of one pig was located within 50 px (a) or 20 px (b) distance from the tail base of another pig.

snout, ears, and tail-related points. Recent advances in pose estimation have highlighted the benefits of using multi-view camera systems or integrating depth sensors to resolve occlusions and obtain more complete 3D representations of body dynamics. Tu et al. [1] demonstrated that pose tracking from lateral camera angles yielded superior temporal continuity and fewer dropped detections. Therefore, future work should evaluate whether combining overhead and lateral views can improve the reliability of behaviourally relevant landmarks, especially for tail-directed interactions.

Thirdly, although visual spray markings supported individual recognition during annotation and video inspection, the present pipeline should not be interpreted as a fully automated long-term re-identification system. Individual identity could be assigned when marking patterns, body position, and spatial continuity were visually clear, but ambiguous cases caused by overlap, occlusion, or poor visibility were treated conservatively. This approach was appropriate for a controlled research setting with a small group of marked pigs, but it is not sufficient for large-scale commercial deployment where animals are typically unmarked and where identity must be maintained continuously across long recording periods. Future systems should therefore integrate pose estimation with dedicated multi-animal tracking, re-identification algorithms, RFID data, or other sensor-based identification methods to improve identity consistency across occlusion events and longer time scales.

Although the analysis of tail-snout proximity events provides a valuable behavioural proxy, the pose estimation models were not explicitly trained or validated on labelled tail-biting or aggressive behaviours. Consequently, the current system cannot directly classify behavioural events but only highlight spatial and temporal conditions under which such interactions may occur. This limitation is increasingly being addressed in recent studies: Souza et al. [33] proposed a hybrid CNN-transformer architecture that classifies pig aggressive behaviour based on pose sequences and spatiotemporal features, showing promising results for multi-individual interaction detection. Similarly, Hakansson and Jensen [2] advocated for combining pose-based features with ethogram-driven annotation to train robust classifiers for tail-biting prediction in group-housed pigs. In line with these recommendations, future work should combine the present pose estimation pipeline with manually annotated tail-biting events, temporal sequence modelling, and supervised behaviour classification to distinguish incidental proximity from true tail-directed biting or manipulation.

Spatial proximity analysis was performed using fixed pixel thresholds (20 px and 50 px) to define interactions between body parts such as tails and snouts. However, these thresholds do not take into account inter-individual differences in body size, body orientation, or image distortion caused by camera lens properties. This introduces potential bias in proximity-based inference. A more robust approach would involve normalizing distances based on estimated pig body length or applying camera calibration parameters to correct for spatial distortion. As recommended by Qi et al. [30], proximity-based behaviour estimation should be complemented by adaptive thresholding and perspective-aware modelling to minimize false positives arising from apparent, but non-interactive co-location. Additionally, Tu et al. [1] highlighted the benefits of trajectory-based disambiguation of social proximity events, especially in large group dynamics where contact may be incidental or non-behavioural in nature.

Another practical limitation concerns the use of spray markings for individual pig identification. While this method is convenient and cost-effective in controlled experimental settings, it is inherently temporary and prone to degradation as the markings fade, smear, or become obscured by dirt. This approach is therefore not feasible for long-term or large-scale deployment in commercial farms, where continuous and reliable identification is required. A more robust alternative involves integrating pose estimation systems with automatic identification through electronic ear tags or computer vision-based re-identification modules. Recent advances in sensor fusion have demonstrated that combining visual tracking with RFID or ear-tag-embedded accelerometers can enable persistent individual recognition and data continuity across sessions [28]. Such integration would facilitate the development of end-to-end systems capable of linking posture dynamics with individualized behavioural profiles, thereby enhancing both traceability and welfare monitoring under real farming conditions.

The configuration that provided the strongest overall pose-estimation performance may not necessarily be optimal for every downstream behavioural application. BODY5P offered the best balance of general localization accuracy, anatomical coverage, and robustness in the present dataset, but tail-biting detection depends particularly on the reliable localization of snout and tail-related landmarks. These landmarks are more prone to occlusion, rapid movement, and ambiguous visibility in group-housed pigs. Therefore, future behaviour-specific pipelines should evaluate keypoint configurations not only using global metrics such as RMSE, *m*AP, and *m*AR, but also using task-specific metrics such as snout and tail localization error, dropout rates of behaviourally relevant landmarks, and agreement with manually annotated tail-directed interactions.

These limitations underscore the importance of continued development in annotated datasets, multi-view and multimodal sensing, automated identity tracking, behaviour-specific model training, adaptive post-processing strategies, and practical edge-level deployment to bridge the gap between experimental pose estimation pipelines and robust, scalable solutions for farm-based animal welfare monitoring.

4. Conclusions

This study provides a comparative evaluation of different keypoint configurations and neural network architectures for pose estimation in the context of pig behaviour monitoring. The results showed that skeletal model design substantially affected pose estimation performance in group-housed pigs. Among the evaluated configurations, BODY5P provided the best overall balance between localization accuracy, anatomical coverage, and robustness, likely because dorsal body landmarks were more consistently visible and less affected by occlusion. However, the most accurate configuration for general pose estimation should not automatically be considered optimal for all downstream behavioural tasks, since tail-directed behaviours depend strongly on reliable snout and tail localization.

The comparison of neural network architectures demonstrated that deep learning models can be effectively applied to complex animal pose estimation tasks with strong accuracy, even on modest hardware. HRNet-W32 achieved the highest overall accuracy and the most stable localization performance, whereas lightweight models such as DLCRNet trained substantially faster and exhibited reliable generalization, reinforcing their potential for portable systems and automated on-farm monitoring setups.

The temporal and spatial analyses showed that pose estimation outputs can provide useful indicators of tracking reliability, space use, and candidate social interactions. However, tail-snout proximity should be interpreted as an exploratory spatial indicator rather than as a validated tail-biting detection system. Future work should focus on expanding the diversity and volume of annotated training data, improving video resolution and temporal sampling, integrating multi-view cameras and automated identity tracking, and combining pose estimation with manually annotated behavioural events to support reliable detection of welfare-relevant behaviours in real-world agricultural settings.

Ethical statement

This study was conducted using video recordings collected as part of routine farm management practices. No additional experimental manipulation, invasive procedures, or handling of animals was performed specifically for the purposes of this research.

All animals were housed and managed in accordance with national and institutional regulations on animal welfare. The use of video-based observation ensured that animal behavior was monitored without causing additional stress or disturbance.

According to local and institutional guidelines, ethical approval was not required for this type of observational study.

CRediT authorship contribution statement

Kirill Ivanov: Writing – original draft, Software, Methodology, Investigation, Formal analysis. **Valentina Bonfatti:** Writing – review & editing, Supervision, Project administration. **Claudia Kasper:** Supervision, Investigation, Data curation. **Hassan-Roland Nasser:** Supervision, Software, Resources, Project administration, Investigation, Funding acquisition, 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.

Acknowledgments

This research was supported by the European Cooperation in Science and Technology (COST) Action CA22112 – European Network for Innovative Phenotyping of Livestock for Improved Animal Health, Welfare, and Productivity (EU-LI-PHE). The author gratefully acknowledges the support received through Virtual Mobility Grant 2024 under the same Action, which facilitated collaboration and data sharing between research teams at the University of Padova (Italy) and Agroscope (Switzerland).

Special thanks are extended to Dr. Hassan-Roland Nasser, Dr. Claudia Kasper-Völkl, Prof. Valentina Bonfatti, and Dr. Kasper Gross for their guidance, discussions, and contributions to the experimental design and data analysis. The study also benefited from the networking and capacity-building activities organized within the EU-LI-PHE COST network, which provided an interdisciplinary platform for developing and disseminating methodologies for behaviour recognition and welfare monitoring in pigs.

We gratefully acknowledge financial support from the Italian Ministry of Health (Ricerca Corrente IZS VE 09/22 RC - *GeneTail* project) and the Department of Comparative Biomedicine and Food Science, University of Padova (Department of Excellence 2023 - SENTINEL project).

Data availability

The data that has been used is confidential.

References

[1] S. Tu, M. Li, X. Zhang, et al., Behaviour tracking and analyses of group-housed pigs based on improved ByteTrack, Animals 14 (22) (2024) 3299, https://doi.org/10.3390/ani14223299.
[2] F. Hakansson, D.B. Jensen, Automatic monitoring and detection of tail-biting behaviour in groups of pigs using video-based deep learning methods, Front. Vet. Sci. 9 (2023) 1099347, https://doi.org/10.3389/fvets.2022.1099347.
[3] A. Mathis, P. Mamidanna, K.M. Cury, T. Abe, V.N. Murthy, M.W. Mathis, M. Bethge, DeepLabCut: markerless pose estimation of user-defined body parts with deep learning, Nat. Neurosci. 21 (9) (2018) 1281–1289, https://doi.org/10.1038/s41593-018-0209-y.
[4] S.J. Devi, et al., Analysis of pig posture detection in group-housed pigs using deep learning-based mask scoring instance segmentation, Anim. Sci. J. 95 (1) (2024) e13975, https://doi.org/10.1111/asj.13975.
[5] T.D. Pereira, D.E. Aldarondo, L. Willmore, M. Kislin, S.S.-H. Wang, M. Murthy, J. W. Shaevitz, Fast animal pose estimation using deep neural networks, Nat. Methods 16 (2019) 117–125, https://doi.org/10.1038/s41592-018-0234-5.
[6] T.M.C.G. de Paula, R.V. de Sousa, M.P. Sarmiento, et al., Deep learning pose detection model for sow locomotion, Sci. Rep. 14 (2024) 16401, https://doi.org/10.1038/s41598-024-62151-7.
[7] Q. Guo, D. Pei, Y. Sun, P.P. Langenhuizen, C.A. Orsini, K.H. Martinsen, P. Bijma, Multi-Object Keypoint Detection and Pose Estimation for Pigs, in: Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications 3, 2025, pp. 466–474.
[8] M. Wutke, et al., Detecting animal contacts—a deep learning-based pig detection and tracking approach for the quantification of social contacts, Sensors 21 (22) (2021) 7512.
[9] C.Q. Liu, H.J. Ye, S.H. Lu, Z. Tang, Z. Bai, L. Diao, Skeleton extraction and pose estimation of piglets using ZS-DLC-PAF, Int. J. Agric. Biol. Eng. 16 (3) (2023) 180–193, https://doi.org/10.25165/j.ijabe.20231603.6930.
[10] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation. CVPR, 2019, pp. 5693–5703, https://doi.org/10.1109/CVPR.2019.00584.
[11] C. Chen, W. Zhu, T. Norton, Behaviour recognition of pigs and cattle: journey from computer vision to deep learning, Comput. Electron. Agric. 187 (2021) 106255, https://doi.org/10.1016/j.compag.2021.106255.
[12] T. Norton, C. Chen, M.L.V. Larsen, D. Berckmans, Precision livestock farming: building 'digital representations' to bring the animals closer to the farmer, Animal. 13 (12) (2019) 3009–3017, https://doi.org/10.1017/S175173111900199X.
[13] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. CVPR 2016, 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.
[14] T. Nath, A. Mathis, A.C. Chen, A. Patel, M. Bethge, M.W. Mathis, Using DeepLabCut for 3D markerless pose estimation across species and behaviors, Nat. Protoc. 14 (2019) 2152–2176, https://doi.org/10.1038/s41596-019-0176-0.
[15] J.M. Graving, D. Chae, H. Naik, L. Li, B. Koger, B.R. Costelloe, I.D. Couzin, DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning, Elife 8 (2019) e47994, https://doi.org/10.7554/eLife.47994.
[16] J. Lauer, M. Zhou, S. Ye, W. Menegas, T. Nath, A. Mathis, Multi-animal pose estimation and tracking with DeepLabCut, Nat. Methods 19 (2022) 496–504, https://doi.org/10.1038/s41592-022-01447-0.
[17] T.D. Pereira, J.W. Shaevitz, M. Murthy, Quantifying behavior to understand the brain, Nat. Neurosci. 25 (2022) 1045–1056, https://doi.org/10.1038/s41593-022-01145-7.
[18] M. Kabra, A.A. Robie, M. Rivera-Alba, S. Branson, K. Branson, JAABA: interactive machine learning for automatic annotation of animal behavior, Nat. Methods 10 (1) (2013) 64–67, https://doi.org/10.1038/nmeth.2281.
[19] M. Kashiha, C. Bahr, S. Ott, C.P. Moons, T.A. Niewold, F.O. Ödberg, D. Berckmans, Automatic identification of marked pigs in a pen using image pattern recognition, Comput. Electron. Agric. 93 (2013) 111–120, https://doi.org/10.1016/j.compag.2013.02.002.
[20] I. Camerlink, S. Menneson, S.P. Turner, M. Farish, G. Arnott, Lateralization influences contest behaviour in domestic pigs, Sci. Rep. 8 (2018) 12116, https://doi.org/10.1038/s41598-018-30634-z.
[21] E.I. Brunberg, et al., Omnivores going astray: a review and new synthesis of abnormal behavior in pigs and laying hens, Front. Veterin. Sci. 3 (2016) 57.
[22] H.A. van de Weerd, J.E. Day, A review of environmental enrichment for pigs housed in intensive housing systems, Appl. Anim. Behav. Sci. 116 (1) (2009) 1–20, https://doi.org/10.1016/j.applanim.2008.08.001.
[23] A. Valros, S. Ahlström, H. Rintala, T. Häkkinen, H. Saloniemi, The prevalence of tail damage in slaughter pigs in Finland and associations to carcass condemnations, Acta Agric. Scandinavica, Sect. A—Anim. Sci. 54 (4) (2004) 213–219, https://doi.org/10.1080/09064700510009234.
[24] M.L.V. Larsen, H.M.L. Andersen, L.J. Pedersen, Can tail damage outbreaks in the pig be predicted by behavioural change? The Veterin. J. 209 (2016) 50–56, https://doi.org/10.1016/j.tvjl.2015.12.001.
[25] K. Luxem, P. Mocellin, F. Fuhrmann, J. Kürsch, S.R. Miller, J.J. Palop, P. Bauer, Identifying behavioral structure from deep variational embeddings of animal motion, Commun. Biol. 5 (1) (2022) 1267, https://doi.org/10.1038/s42003-022-04080-7.
[26] M.W. Mathis, A. Mathis, Deep learning tools for the measurement of animal behavior in neuroscience, Curr. Opin. Neurobiol. 60 (2020) 1–11, https://doi.org/10.1016/j.conb.2019.10.008.
[27] H. Gan, M. Ou, E. Huang, C. Xu, S. Li, J. Li, Y. Xue, Automated detection and analysis of social behaviors among preweaning piglets using key point-based spatial and temporal features, Comput. Electron. Agric. 188 (2021) 106357, https://doi.org/10.1016/j.compag.2021.106357.
[28] Y. Huang, Y. Lv, D. Xiao, J. Liu, Z. Tan, G. Li, Behavior classification and rhythm analysis of pigs based on wearable sensors, Smart Agric. Technol. (2025) 101457, https://doi.org/10.1016/j.atech.2025.101457.
[29] Y. Pu, Y. Zhao, H. Ma, J. Wang, A lightweight pig aggressive behavior recognition model by effective integration of spatio-temporal features, Animals, 15 (8) (2025) 1159, https://doi.org/10.3390/ani15081159.
[30] Qi, F., Hou, Z., Lin, E. et al. (2025). Pig behaviour dataset and spatial-temporal perception and enhancement networks based on the attention mechanism for pig behaviour recognition. arXiv preprint 2503.09378. https://doi.org/10.48550/arXiv.2503.09378.

[31] J.-H. Witte, P. Heseker, J. Probst, et al., Image-based tail posture monitoring of pigs, in: Proc. 57th Hawaii International Conference on System Sciences (HICSS-57), 2024. https://hdl.handle.net/10125/107327.
[32] C. Jeon, H. Kim, D. Kim, A deep-learning-based system for pig posture classification: enhancing Sustainable Smart Pigsty Management, Sustainability. 16 (7) (2024) 2888, https://doi.org/10.3390/su16072888.
[33] Souza, J.S., Bedin, E., Higa, G.T.H., Loebens, N., & Pistori, H. (2024). Pig aggression classification using CNN, transformers and recurrent networks. *arXiv preprint arXiv:2403.08528*. https://doi.org/10.48550/arXiv.2403.08528.