Bachelor Thesis

Markerless Pose Estimation and Mice Tracking with Deep Learning

Technische Universität Berlin

Mohammad Orabe

September 1, 2022

Reviewer:
Prof. Dr. York Winter
Prof. Dr. rer. nat. Klaus Obermayer

Technische Universität Berlin
Fakultät IV Elektrotechnik und Informatik
https://www.eecs.tu-berlin.de/menue/fakultaet_iv

I hereby declare that the thesis submitted is my own, unaided work, completed without any unpermitted external help. Only the sources and resources listed were used.

Berlin, September 1, 2022
Mohammad Orabe

Abstract

Analyzing social interaction dynamics in multi-agent environments requires reliable localization, tracking, and interpretation of individual behaviors over time. Automated pose estimation and identity-preserving tracking are therefore essential for quantitatively studying complex interaction patterns, particularly under conditions involving occlusions, rapid motion, and visually similar subjects.

This thesis investigates a deep learning-based approach for markerless pose estimation and behavioral analysis using transfer learning and fine-tuning strategies. A convolutional neural network with a ResNet-based backbone pretrained on large-scale image datasets was adapted to the task by fine-tuning on a comparatively small set of manually annotated video frames. The model learns spatial representations of anatomically relevant body landmarks directly from raw video data, enabling robust localization and tracking across diverse recording conditions.

Beyond model adaptation, the work presents the design and implementation of a complete experimental pipeline, covering data acquisition, frame extraction, annotation, model training, inference, and evaluation. To improve identity consistency in multi-animal scenarios, RFID sensing was incorporated as an auxiliary modality providing persistent identity cues. A synchronization and matching strategy was developed to associate network predictions with RFID detections, complemented by a dedicated Correction of Switched Identities (CSI) mechanism to resolve tracking inconsistencies arising from occlusions and identity swaps. The resulting system supports end-to-end processing, including downstream behavioral analysis and classification.

Experimental evaluation demonstrates that the fine-tuned deep neural network achieves high-precision pose estimation despite limited training data. The final model achieved a test error of 4.9 pixels, comparable to inter-annotator labeling variability (5.2 pixels), indicating near human-level prediction accuracy.

Zusammenfassung

Die Analyse sozialer Interaktionsdynamiken in Multi-Agenten-Umgebungen erfordert eine zuverlässige Lokalisierung, Verfolgung und Interpretation individueller Verhaltensmuster über die Zeit. Automatisierte Posenschätzung und identitätskonsistente Verfolgung stellen daher zentrale Voraussetzungen für die quantitative Untersuchung komplexer Interaktionsprozesse dar, insbesondere unter Bedingungen mit Verdeckungen, schnellen Bewegungen und visuell schwer unterscheidbaren Individuen.

Diese Arbeit untersucht einen Deep-Learning-basierten Ansatz zur markerlosen Posenschätzung und Verhaltensanalyse unter Verwendung von Transfer-Learning- und Fine-Tuning-Strategien. Hierzu wurde ein konvolutionales neuronales Netzwerk mit einer ResNet-basierten Architektur, das auf großskaligen Bilddatensätzen vortrainiert wurde, durch Fine-Tuning an die spezifische Aufgabenstellung angepasst. Das Modell wird auf einer vergleichsweise kleinen Menge manuell annotierter Videoframes trainiert und lernt räumliche Repräsentationen anatomisch relevanter Körperlandmarken direkt aus den Rohvideodaten. Dadurch wird eine robuste Lokalisierung und Verfolgung über unterschiedliche Aufnahmebedingungen hinweg ermöglicht.

Über die reine Modellanpassung hinaus umfasst die Arbeit die Konzeption und Implementierung einer vollständigen experimentellen Pipeline, einschließlich Datenakquisition, Frame-Extraktion, Annotation, Modelltraining, Inferenz sowie systematischer Evaluierung. Zur Verbesserung der Identitätskonsistenz in Multi-Tier-Szenarien wurde RFID-Sensorik als ergänzende Modalität integriert, die persistente Identitätsinformationen bereitstellt. Eine Synchronisations- und Matching-Strategie wurde entwickelt, um Netzwerkausgaben mit RFID-Detektionen räumlich und zeitlich zu verknüpfen. Zusätzlich wurde ein Verfahren zur Korrektur vertauschter Identitäten (Correction of Switched Identities, CSI) eingeführt, um Tracking-Fehler infolge von Verdeckungen und Identitätswechseln zu erkennen und zu beheben. Das resultierende System unterstützt eine durchgängige Verarbeitungskette bis hin zur Verhaltensanalyse und Klassifikation.

Die experimentellen Ergebnisse zeigen, dass das feinjustierte neuronale Netzwerk trotz begrenzter Trainingsdaten eine hohe Posenschätzgenauigkeit erreicht. Das finale Modell erzielt einen Testfehler von 4,9 Pixeln, was der Variabilität menschlicher Annotatoren (5,2 Pixel) entspricht und somit eine nahezu menschliche Vorhersagegenauigkeit indiziert.

Contents

List of Figures

List of Tables

1 Introduction

Locating and tracking moving objects or organisms in spatiotemporal data constitutes a fundamental problem with broad relevance across numerous scientific disciplines [6]. Accurate motion and pose analysis underpin a wide range of applications, including behavioral analysis, biomechanics, robotics, and neuroscience. In particular, neuroscience increasingly relies on quantitative behavioral measurements to investigate motor function, cognition, and social interaction. For instance, gait analysis provides valuable markers for characterizing and diagnosing movement disorders such as Parkinson's disease [14], while the systematic study of social interactions offers insight into both typical and pathological cognitive and emotional processes [4]. Many of these research paradigms depend critically on precise estimation of subject positions, trajectories, and postures from recorded video data.

Historically, behavioral quantification has often relied on manual annotation, whereby human observers identify and track points of interest across consecutive video frames. Although conceptually simple, manual labeling is inherently tedious, time-intensive, and prone to variability arising from observer fatigue and subjective bias. These limitations become particularly pronounced in experiments involving multiple individuals, rapid movements, or frequent interactions, where reliable frame-by-frame annotation becomes impractical.

Computer vision methods provide a scalable alternative by automating key aspects of localization and tracking. Traditional approaches frequently employ physical markers attached to specific body parts, enabling precise pose estimation and motion tracking. While marker-based systems can achieve high spatial accuracy [9], their use may interfere with natural behavior, potentially altering experimental outcomes. Furthermore, marker placement is labor-intensive and may not be feasible in all experimental contexts. These challenges have motivated the growing adoption of markerless tracking techniques that infer body part locations directly from visual data.

Recent advances in deep learning have significantly transformed markerless pose estimation and tracking methodologies. Deep neural networks enable data-driven representation learning, allowing models to extract complex visual features and motion patterns directly from raw video frames [11, 5]. Unlike rule-based systems, deep learning approaches can capture highly nonlinear relationships and demonstrate strong robustness to noise, variability, and partial occlusions. Large-scale empirical successes across computer vision tasks further highlight their effectiveness [10, 8]. In comparison to earlier techniques, deep learning-based methods typically offer improved accuracy, scalability,

and generalization performance [2].

These methodological developments are particularly relevant for the study of social behavior. Social interactions inherently involve the dynamic coordination of actions between multiple individuals, often characterized by rapid and diverse behavioral transitions. In rodents, and especially in mice, social behavior reflects complex internal states, including motivation and affect, as well as responses to conspecific actions. Recent studies have further revealed that neural dynamics may exhibit inter-brain coupling during social engagement, despite the absence of direct physical connections between interacting brains. Such findings emphasize the importance of precise behavioral tracking for investigating the neural mechanisms underlying social cognition and interaction.

Quantitative analysis of social behavior, however, poses distinct computational challenges. Video-based tracking systems must reliably localize multiple anatomically relevant body parts while simultaneously maintaining consistent identities for visually indistinguishable subjects. Identity preservation becomes particularly difficult in scenarios involving close interactions, overlapping body configurations, and similar visual appearance. Purely vision-based tracking methods may therefore produce identity ambiguities or swaps, which can substantially compromise downstream behavioral analyses.

To address these challenges, deep learning-based markerless tracking frameworks have emerged as powerful tools. Among these, DeepLabCut (DLC) [13] has become a widely adopted solution for animal pose estimation, enabling supervised deep neural networks to predict body part locations directly from raw video data. By leveraging learned spatial representations, DLC provides robust posture estimation without requiring intrusive physical markers.

Building upon these advances, the objective of this thesis is to evaluate the viability of a deep learning-based approach for pose estimation and multi-animal tracking of mice across distinct experimental paradigms. Particular emphasis is placed on resolving the identity preservation problem in multi-individual recordings. To this end, we develop a fully automated Python-based pipeline that integrates markerless pose estimation via DeepLabCut with RFID-based identity information.

The proposed system is organized into two principal stages. **Stage I** performs pose estimation and multi-animal tracking using DLC, during which a deep neural network model is created, trained, and evaluated to predict body part locations and trajectories. **Stage II** incorporates RFID detections to replace temporary tracking identities with persistent RFID-derived labels, thereby resolving visual identity ambiguities.

To further enhance robustness, the pipeline includes dedicated validation and correction mechanisms that monitor tracking consistency and detect potential identity swaps. When ambiguities are identified, automated re-matching procedures are applied to restore correct identity assignments. Collectively, this framework enables reliable body part tracking and identity preservation while relying exclusively on raw video data and non-visual identity cues.

All experiments were conducted in the Winter laboratory at Humboldt University.

2 Theoretical Background

2.1 Artificial Neural Networks

Artificial Neural Network (ANN) is a method in machine learning used to understand complex data by learning intricate structures in high-dimensional data [12] . The ANN is inspired by the biological neural networks of the animal brains. The basic computational unit of an ANN, the artificial neuron, in a similar fashion, loosely models the biological neurons that make up an animal brain. Each neuron consists of weighted incoming connections from other neurons and a bias. These incoming connections are integrated by summing all the weights and adding the bias of the current neuron. The output of a neuron is the result of applying an activation function on the aforementioned weighted sum, which is then passed to neurons in the next layer via its outgoing connections.

2.2 Optimizing Artificial Neural Networks

The process of calculating the output of neurons from the first layer, the input layer, and all the way to the last output layer, is called forward propagation. At the end of this process, a loss term is calculated to measure the difference between the output of the ANN and the model's labels using a loss function. The aim of the optimization is to minimize this loss term. This is done by updating every parameter for each neuron going from the last layer to the first input layer based on the loss term, also called the backward propagation.

2.3 Convolutional Neural Networks (CNN)

Convolutional Neural Network (CNN) is a class of ANN that is commonly used in image pattern recognition [1] . It takes its name from the mathematical operation between matrices called convolution. The structure of a typical CNN can be regarded as a series of stages. The first stages contain two type of layers: convolutional layers and pooling layers, whereas the last stages are composed of fully connected layers. The convolutional and fully-connected layers have parameters, whereas the pooling don't have any [12] . Units in a convolutional layer are organized in feature maps, which partially overlap so that they represent the entire visual field. Each unit receives input from local patches of

feature maps of the previous layer. The pooling layers aims to combine similar features and at the same time reduce the dimensions of data. This is done by combining the outputs of clusters of neurons from the previous layer into one single neuron. The fully connected layers usually come last in the architecture, in which every neuron of the previous layer is connected to every neuron of the next layer.

2.4 ResNet

The deeper the CNN increases, the more complex it becomes to train, and the higher the training error and hence the test error becomes, which negatively affects the performance of the network. In residual networks (ResNets) introduced by He et.al., instead of mapping the input x directly to a function y = f(x), the layer is parametrized to learn a residual mapping y = x + f(x) using skip connections.

The advantage of the residual mapping is that it improves optimization and regularization, allowing us to stack more layers on top of each other without adding more complexity, and thus create very deep neural networks that can solve more complex problems. [7]

2.5 Transfer Learning

Transfer learning is a machine learning method, where we use parameters from a network that has been trained on one task to perform another related task by transferring information. It is based on using the knowledge a model has learned from a task with a lot of available labeled training data in a new task that doesn't have much data. Instead of starting the learning process from scratch, we start with patterns learned from solving a related task. The advantage is that these networks are pretrained on larger datasets so that when it comes to a situation when not much data or time are available it can be very sufficient as it decreases the training time for a neural network model and the data number, resulting in lower generalization error which means eventually better performance [15].

3 Materials and Methods

3.1 Experimental Setup

The experiment was conducted in a custom mouse home-cage environment housing RFID-implanted mice. A 5-megapixel Raspberry Pi camera was mounted above the cage to continuously record animal behavior. The camera module was equipped with two infrared LEDs, enabling automatic switching between day and night acquisition modes. While the implementation utilized a Raspberry Pi camera, the setup is compatible with any camera providing comparable resolution and infrared capability. The camera was interfaced with a Raspberry Pi running the Raspbian operating system for continuous video capture.

To obtain identity information independent of visual tracking, an RFID reader device was installed beneath the customized home-cage. The RFID system was connected to a Windows-based workstation via an Ethernet interface, enabling real-time acquisition of tag detections. The combined hardware configuration is illustrated in Figures 3.1 and 3.2.

3.2 Animals

Animals were housed in standard plastic cages resembling conventional home-cage environments, equipped with bedding material, ad libitum access to water via mounted bottles, and food pellets attached to the cage walls. Prior to and throughout the recording sessions, mice were maintained in an animal-friendly laboratory environment under controlled temperature and humidity conditions. A 12-hour light–dark cycle was employed to ensure stable circadian conditions, enabling behavioral recordings during both the dark phase, when mice are typically most active, and the light phase, when resting behavior predominates.

A total of nine female laboratory mice were included in the experiment. To introduce controlled visual variability, animals differed in body size and fur coloration, comprising three mice each with black, gray, and white coats. This variation was intentionally introduced to improve the robustness and generalization capability of the tracking model under diverse visual conditions. Prior to the experiment, mice were implanted with RFID glass tags to enable non-visual detection and identity tracking via RFID sen-

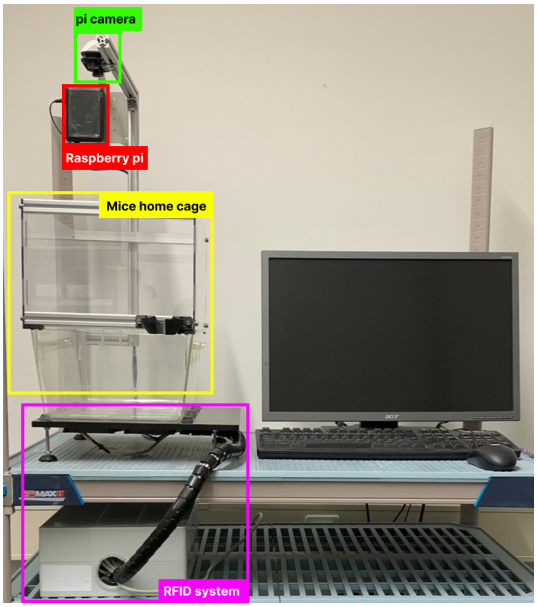


Figure 3.1: Overview of the physical experimental setup. The figure illustrates the customized mouse home-cage equipped with an overhead Raspberry Pi camera for continuous video acquisition. Infrared illumination enables reliable recording under varying lighting conditions. An RFID reader system is installed beneath the cage to capture non-visual identity information, while the Raspberry Pi handles video streaming and data acquisition.

sors. For technical validation of the identification pipeline, RFID capsules were applied exclusively to the black-coated mice.

Post-recording inspection revealed two conditions of particular relevance for model performance. First, mice frequently engaged in close physical interactions, leading to persistent occlusions and overlapping body configurations, which increase tracking ambiguity and necessitate larger training datasets. Second, mice with white and gray fur exhibited reduced visual contrast against the bedding material, complicating reliable body part detection and contributing to increased occlusion-related tracking errors. These observations highlight intrinsic challenges of markerless multi-animal tracking in naturalistic home-cage environments.

3.3 RFID Tagging

The mice used in the home-cage environment exhibit highly similar visual appearance, which poses a fundamental challenge for reliable identity preservation in multi-animal tracking. Although the DeepLabCut (DLC) framework enables accurate markerless pose

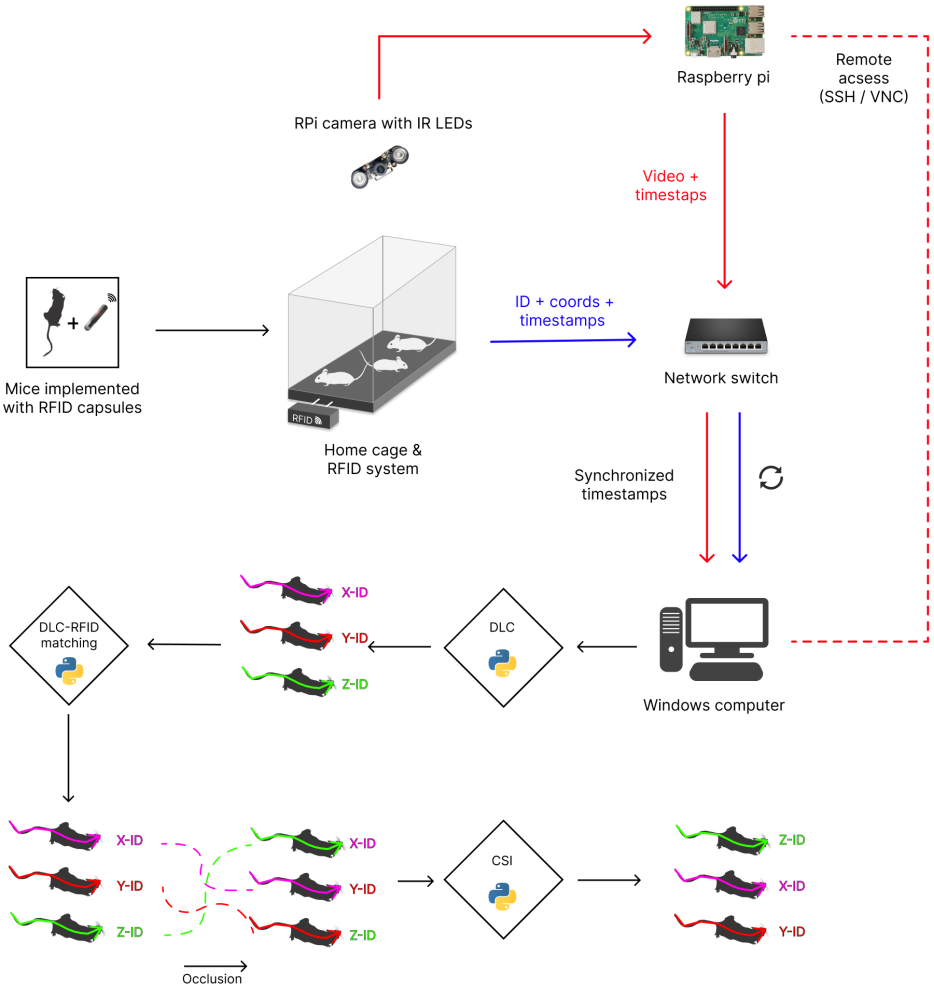


Figure 3.2: Abstract representation of the experimental pipeline and multimodal data integration framework. Video data are acquired using an overhead Raspberry Pi camera and transmitted to a Windows workstation via a Raspberry Pi device. In parallel, RFID sensors positioned beneath the home-cage generate time-stamped identity detections. Both data streams are temporally synchronized to enable cross-modal association. DeepLabCut (DLC) performs markerless pose estimation and multi-animal tracking, producing body-part trajectories with temporary identities. These identities are subsequently resolved through DLC–RFID matching based on spatial overlap within reader regions. The Correction of Switched Identities (CSI) module further refines identity consistency by detecting and correcting potential tracking ambiguities arising from occlusions and identity swaps.

estimation and tracking, the algorithm assigns temporary or dummy identities to individual animals. In scenarios involving close interactions, rapid movements, or partial occlusions, vision-based tracking alone may therefore produce identity ambiguities or unintended identity switches.

To mitigate this limitation, RFID technology was incorporated as an independent source of persistent identity information. Following pose estimation and tracking, RFID detections were used to validate the analyzed video data and to associate each tracked animal with its unique tag-derived identity. The RFID system consisted of eight spatially distributed readers positioned beneath the home-cage. During recording, the RFID device cyclically scanned all readers at 200 ms intervals. Whenever a tag entered the detection range of a reader, a new event was registered and initialized with a counter value. In subsequent scanning cycles, the counter was incremented if the same tag remained within range (Figure 3.4).

Due to the spatial proximity of readers, a single tag could occasionally be detected by multiple readers simultaneously, introducing positional uncertainty and measurement noise. Nevertheless, RFID detections provided robust identity information that is invariant to visual occlusions. In the final processing stage, RFID-derived identities replaced the temporary DLC tracking labels, enabling consistent subject identification. Additionally, the pipeline included validation and correction mechanisms designed to detect potential identity swaps and perform automated re-matching when inconsistencies were observed.

3.4 Data Acquisition

Communication between the Raspberry Pi (RPi) and the Windows workstation was established using the Secure Shell (SSH) protocol, enabling remote control and reliable data exchange. Video data were captured on the RPi and encoded using the H.264 compression standard. To facilitate subsequent multimodal synchronization, the timestamp corresponding to the first acquired video frame was stored separately in a CSV file. This reference timestamp served as the temporal anchor for aligning video recordings with RFID detections during the identity matching stage.

Continuous video streaming was implemented via a lightweight network-based transmission protocol. The system comprised two dedicated scripts: a server application executed on the Windows machine, responsible for listening for incoming connections, and a client application running on the RPi, which transmitted a continuous sequence of image frames. Video recordings were acquired at a spatial resolution of 1280 × 720 pixels and a frame rate of 25 frames per second (FPS).

To capture a broad range of naturalistic behaviors and postural variations, three independent 24-hour recording sessions were conducted. Each session included three mice of the same strain and fur coloration housed within the home-cage environment. This

design ensured controlled visual conditions while preserving behavioral diversity across recordings.

RFID data were acquired independently on the Windows workstation. The RFID system generated time-stamped detection events stored in CSV format, where each entry contained the detected tag identity, the corresponding reader location, and the detection timestamp. These RFID measurements provided persistent identity information that was later integrated with the video-based tracking results.

3.5 Frame Extraction

To construct a representative training dataset, the three continuous recording sessions were partitioned into shorter video segments. This segmentation strategy enabled systematic sampling across varying behavioral states, postural configurations, and lighting conditions. Frames were subsequently extracted from these video snippets to capture the diversity of observed animal behaviors.

A total of 288 frames were selected for model training. Frame selection was performed iteratively to progressively refine dataset quality and improve detector robustness. In the initial stage, frames were sampled using a k-means clustering strategy applied to the full set of 24-hour recordings. This unsupervised procedure promoted the selection of visually diverse samples, thereby reducing redundancy and ensuring broad coverage of behavioral and environmental variability.

Following preliminary model evaluation, additional frames were incorporated through targeted manual sampling. Specifically, frames depicting close physical interactions, partial occlusions, and failure cases identified during visual inspection were preferentially selected. This refinement step aimed to improve model performance under challenging conditions, particularly those involving overlapping animals and reduced visual separability. A summary of the frame extraction procedure is provided in Table **??**.

3.6 Frame Annotation

Manual annotation constitutes a critical component of supervised pose estimation and represents a major source of effort in dataset preparation, as summarized in Table **??**. The annotation process becomes particularly demanding in multi-animal scenarios involving close interactions and occlusions, where body part visibility may be limited.

All extracted frames were annotated using the graphical user interface provided by the napari plugin. For each animal, twelve anatomically defined keypoints were labeled: snout, left ear, right ear, shoulder, four intermediate spine points, tail base, two tail points, and tail end. These landmarks were selected to capture both head orientation and full-body posture dynamics.

To ensure labeling consistency and accuracy, the built-in label validation utilities of DeepLabCut were employed. This validation procedure enabled visual inspection of annotated frames and facilitated the correction of potential inconsistencies or misplaced keypoints prior to model training.

Iteration Round	Black Mice	Gray Mice	White Mice
1	4	4	4
2	8	8	8
3	24	24	24
4	28	28	28
5	32	32	32
6	20	20	20
7	20	44	39
8	19	42	23
Total Labeled Frames	155	202	178
Total Duration (hours)	8	10	9

Table 3.1: Summary of the manually annotated dataset used for model training and refinement. The table reports the number of labeled frames per iteration round and mouse cohort, along with the total annotation effort measured in labeled frames and recording duration.

3.7 Outlier Extraction and Label Refinement

To improve network robustness under challenging and previously unseen conditions, additional frames associated with elevated prediction errors were identified and incorporated into the training dataset (Table 3.1). These frames predominantly corresponded to failure cases, including partial occlusions, ambiguous body configurations, and reduced visual contrast, where the initial model exhibited degraded labeling accuracy.

DeepLabCut provides dedicated diagnostic and evaluation utilities that enable systematic assessment of tracking quality and detection of outlier frames. These procedures facilitate the identification of samples in which network predictions deviate from plausible anatomical configurations. Frames flagged as outliers were manually inspected and corrected by refining the associated keypoint annotations.

This iterative refinement strategy improves annotation consistency, reduces prediction uncertainty, and enhances model generalization by explicitly exposing the network to visually challenging scenarios encountered during recording.

3.8 Identity Matching Using RFID

In this stage, the temporary identities produced by DeepLabCut (DLC) were resolved by associating each tracked animal with its corresponding RFID tag. While DLC provides reliable pose estimation and multi-animal tracking, identity preservation may degrade during close interactions or occlusions. RFID sensing therefore serves as an independent modality for persistent subject identification.

RFID detections and DLC predictions were temporally synchronized using RFID event timestamps and the reference timestamp of the first acquired video frame. To establish spatial correspondence, the physical locations of RFID readers were mapped onto pixel coordinates within the video frames. Based on this mapping, each frame was partitioned into regions corresponding to the effective coverage areas of individual RFID readers (Figure 3.4).

For each video frame, the spatial agreement between estimated RFID tag positions and DLC-predicted mouse locations was evaluated. Overlap statistics were accumulated across frames, and each RFID identity was assigned to the animal exhibiting the highest spatial correspondence.

3.9 RFID Positional Estimation

RFID measurements are inherently noisy due to overlapping reader coverage and signal propagation effects. A single tag may therefore be detected by multiple readers within a short temporal window, resulting in ambiguous position estimates. Stable identity matching, however, requires a unique positional representation.

To mitigate positional variability, RFID detections were aggregated within fixed temporal intervals. Each interval could contain multiple detection events for a given tag. Rather than treating individual reads independently, a single representative tag position was estimated per interval. If no detections occurred, no position was assigned.

The representative position P_i of a tag within interval i was computed as:

$$P_i = \sum_{e=1}^{n} t_i(e) \cdot f_i(e) \cdot p_i(e)$$

where $p_i(e)$ denotes the two-dimensional pixel location of event e, $t_i(e)$ represents the effective detection duration, and $f_i(e)$ corresponds to the detection frequency within the interval. This weighted formulation stabilizes positional estimates and reduces the influence of transient detections.

Multiple interval lengths (0.5 s, 1 s, 2 s, and 3 s) were evaluated, and the duration yielding the most consistent identity matching performance (Figure 4.7) was selected for the final implementation.

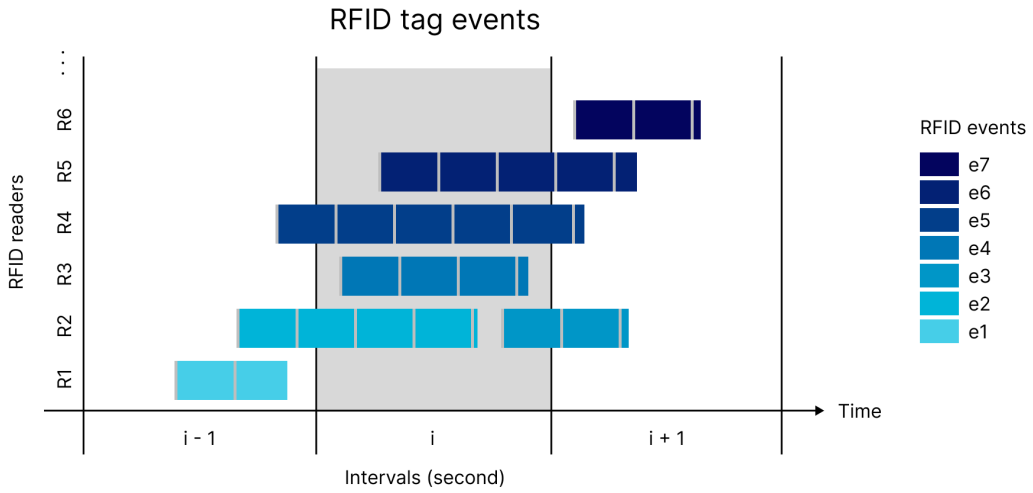


Figure 3.3: Illustration of RFID detection events across time. Each colored bar represents a tag detection associated with a specific RFID reader location (vertical axis). Fixed temporal intervals (horizontal axis) are used to aggregate detections and compute a stable representative tag position within each interval. This aggregation mitigates positional fluctuations arising from multiple-reader detections.

DeepLabCut predictions consist of multiple keypoints per animal. For identity matching, each animal must be represented by a single spatial descriptor. To this end, a centroid was computed as the mean position of stable body landmarks, including the shoulder, spine points (spine1–spine4), and tail base. An overlap within a frame was defined when both the interval-averaged RFID tag position and the DLC-derived centroid fell within the same RFID reader region (Figure 3.4). This criterion enabled robust cross-modal association between visual tracking outputs and RFID-based identity information.

3.10 Correction of Switched Identities (CSI)

Identity swapping may occur when predicted body parts are assembled into continuous tracks across video frames. Although DeepLabCut (DLC) performs joint pose estimation and tracking, temporary inaccuracies in keypoint detection or pose estimation can lead to incorrectly stitched tracklets. In particular, short sequences of degraded predictions may cause misalignment of tracks, resulting in unintended identity switches. DLC assumes that identity consistency is preserved when body part localization remains sufficiently accurate across frames.

In practice, identity disruptions are more likely during inference when video recordings deviate from the visual conditions represented in the training dataset. Variations

in illumination, occlusions, or atypical body configurations may introduce temporary tracking interruptions. To address these issues, a post-processing correction strategy termed *Correction of Switched Identities (CSI)* was developed. The CSI mechanism operates after the ID-matching stage, when each tracked animal has been associated with a persistent RFID-derived identity. The procedure is optional and designed to improve identity consistency during inference.

The CSI approach relies on two complementary mechanisms. First, frames are evaluated sequentially to assess spatial agreement between RFID-derived tag positions and DLC-based mouse detections within the vicinity of RFID reader regions. Second, potential identity switches are detected by monitoring inter-frame displacement of DLC predictions. Specifically, the Euclidean distance between the centroid positions of an individual mouse in consecutive frames is computed. Under normal tracking conditions, centroid displacement remains small due to the high temporal resolution of video acquisition. Empirically, centroid variations of approximately 10 pixels were observed in the absence of identity switches. Frames exhibiting unusually large displacements were therefore treated as candidates for identity inconsistencies.

A displacement threshold was introduced to selectively trigger correction procedures. When the centroid distance exceeded a predefined threshold (100 pixels), the CSI mechanism was activated for the affected frames. Importantly, only frames flagged by this criterion were subjected to correction, rather than the full video sequence. This selective strategy reduces unnecessary corrections and mitigates false positive adjustments.

Identity correction decisions were formulated based on three principal scenarios. In the first scenario, both an RFID tag detection and a DLC-predicted mouse were present within the same reader region, and their spatial positions overlapped. This configuration indicates correct identity assignment, and no correction was applied. In the second scenario, exactly one RFID tag and one DLC-predicted mouse were present within a reader region, but no spatial overlap was observed. In this case, the visual identity was considered inconsistent and reassigned to match the detected RFID tag. In the third scenario, multiple RFID tags or multiple DLC-predicted mice were detected within the same reader region. Due to inherent ambiguity, no automatic correction was performed to avoid erroneous identity modifications.

To determine spatial overlap between RFID and DLC detections, animals were represented by a single positional descriptor. The CSI framework supports either centroid-based representations or the use of selected stable body landmarks. This flexibility allows robust identity verification while accounting for variability in pose estimation accuracy.

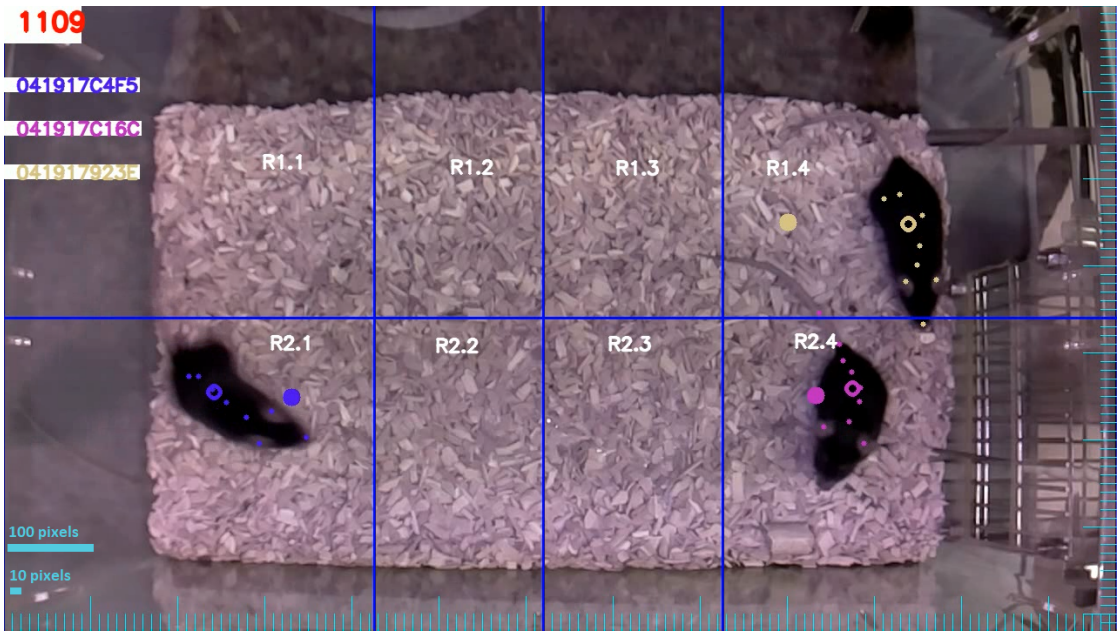


Figure 3.4: Representative video frame after RFID-based identity assignment. Colors denote individual animals. Small markers indicate DeepLabCut-predicted keypoints, circular rings represent centroids computed from stable body landmarks, and large circles correspond to interval-averaged RFID tag detections. The overlaid grid partitions the frame into predefined RFID reader regions, enabling spatial correspondence analysis between vision-based pose estimation and RFID-derived identity information.

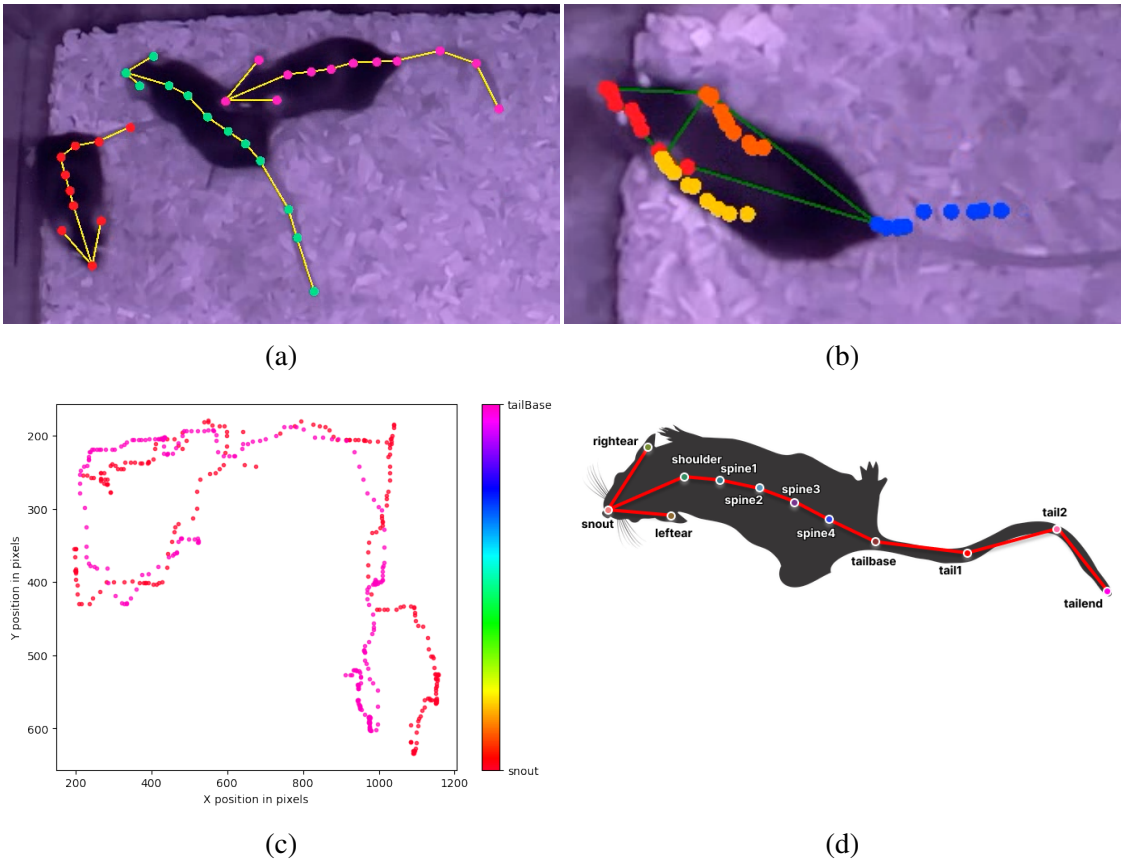


Figure 3.5: Representative DeepLabCut outputs obtained from the multi-animal tracking model. Panels (a) and (b) illustrate pose estimation results under different visualization configurations, highlighting color-coded keypoints and skeletal structures used to represent body-part relationships. Panel (c) shows a trajectory map of selected anatomical landmarks (snout and tail base) across 150 consecutive frames, demonstrating temporal tracking consistency. Panel (d) presents the annotated body-part model defining the keypoints and skeletal connectivity employed for pose estimation.

4 Results

4.1 Performance evaluation

We used a cloud GPU (NVIDIA Quadro A4000, 16 GB GPU, 45 GB RAM) to train our model [add reference: Paperspace gradient]. We accomplished eight consecutive rounds of training. In total, we trained the model for 230,000 iterations, consuming 42 hours of GPU wall time. DeepLabCut provides several loss functions [add Ref. here!] to measure the training progress. A loss function indicates how much the predicted data deviate from ground truth values. To determine whether the function converges to a reliable local minimum, we used the TensorBoard tool to visualize the total loss during the training (Fig. 4.1), where the total loss function is the sum of the cross-entropy loss and the locref loss [add Ref. here!].

While training, we kept an eye on the loss function plateau and periodically stored weights snapshots (checkpoints of the trained model) every 10K iterations. After 6 hours of training ($\approx$ 5000 iterations), the training was stopped automatically to determine if the model's loss as low as possible, to prevent overfitting and ensure that the optimal model is chosen. The snapshots allowed us to resume training from a specific snapshot, with no need to re-train from the beginning after an interruption. However, the most recent snapshot saved during training might not always produce the best performance. Therefore, before expanding the training dataset and continuing training, we evaluated all stored snapshots and chose the one with the best performance.

To evaluate the performance of the trained network, DeepLabCut calculates the Root Mean Square Error (RMSE) between the network's predicted labels and their associated ground truth for each frame and keypoint. We evaluated the model on one shuffle (split) of training and test data, and the test error was 4.9 pixels, while the train error was 3.2 pixels on an image size of 1280 X 720 (fig 4.2). We evaluated the model in each training round, then we expanded our training set, as described in the materials and methods section, and then resumed training (Figure 4.1). Overall, we have manually annotated 535 frames (Table 3.1) containing three mice with the same fur color within an image but with different colors across video frames. We manually annotated 15,900 key points in total (Fig. 4.3). We evaluated several snapshots of trained weights before settling on the training iteration 240K for the final model, as it provides the best accuracy performance on the test dataset. We used the default confidence limit P-cutoff with 0.9 to evaluate the prediction accuracy, where P-cutoff is the prediction likelihood of DeepLabCut. The

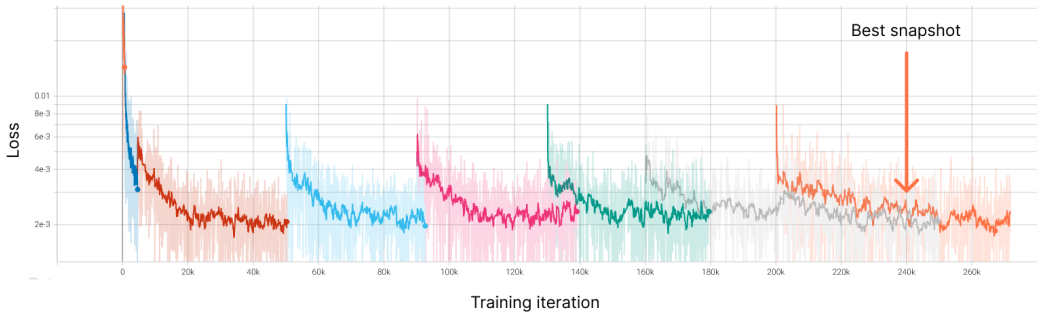


Figure 4.1: Total loss function for ResNet50 when training with when training with the iteratively expanded images. Each colored curve indicates the loss values of a new training round with an expanded training dataset. In each round, training was performed until the loss function converged to a local optimum minimum. 8 training rounds were accomplished in total.

model is more confident in its detections when the likelihood value is high.

Furthermore, we calculated the RMSE between the labels that were predicted by the network and the manually annotated labels both on training and test set. The error was 3.48 pixels for a confidence threshold of 0.9. We also determined the RMSE for each body part. Overall, 75% and 50% of the detection errors were below 4.47 and 3.1 pixels, respectively. The lower the RMSE is, the better is the performance of the model. Figure 4.4) depicts the distribution of these data.

To provide an insight into how precise human annotators are and to measure the variability of the placement of the same labels on the same images that are annotated by two different humans, we evaluated the labeling performance on 20 images (containing three animals) that we had used in the model dataset. An inexperienced student was instructed to annotate approximately 720 body parts. We used the RMSE metric to measure the variability of the annotator's labels. The *tailbase* performed the best with 2.8 pixels and while the *spine* labels had the lowest accuracy (6-7 pixels). The average RMSE across all body parts was 5.2 pixels. Figure 4.5 contrasts and compares the labeling accuracy between two human annotators, as well as with that network's predictions.

4.2 Evaluation with novel videos

We assessed the tracking performance with the Multi-Object Tracking Accuracy metric (MOTA) [3]. All potential errors were considered: number of False Positive (FP) predictions, number of False Negative (FN) predictions and the number of mismatched identities (ID_error). We used the following formula to calculate the MOTA value:

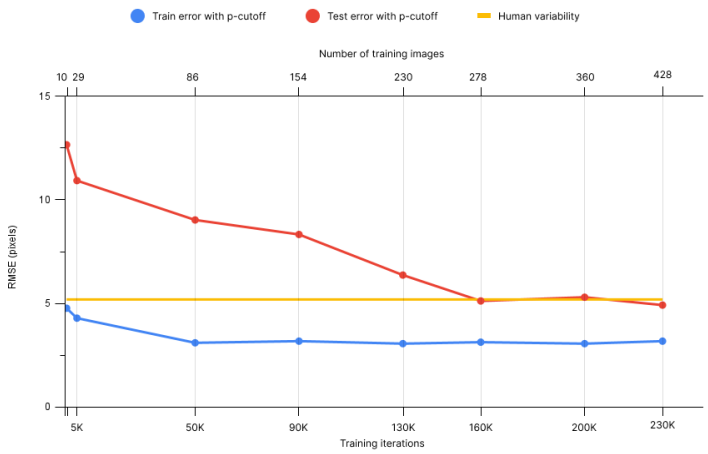


Figure 4.2: Mean average euclidean error (RMSE) between the network's predicted labels and their nearest ground truth neighbors (the manually annotated labels) on training and test images on a split of 80% and 20%, as well as between two human annotators. The prediction accuracy is measured using a P-cutoff of 0.9. Human variability is represented by an orange line, while blue and red curves indicate train error and test error, respectively. Dots indicate the number of training iterations and the number of training images used in a training round. A training round is evaluated every ≈ 50,000 iterations and expanded with new training images. The network achieved human-level accuracy on the test set.

$$MOTA = 1 - \frac{\sum_f (FN_f + FP_f + ID\_error_f)}{\sum_f N_f}$$

where f is the corresponding frame number. FN_f is the total number of missing mice where they are present in the ground truth but not got detected. FP_f is the total number of predicted mice, but they are not present in the ground truth. ID_error_f is the total number of mismatched identities. N_f is a total number of mice present in the ground truth.

Moreover, we used the Mean Average Precision (mAP) metric to provide further prediction evaluation. We calculated the Precision and Recall sub-metrics, which are defined as follows:

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

We established three different datasets (Vid-10, VidFewIDSwaps and VidManyIDSwaps) containing novel videos with 1280x720 pixels resolution, each with three mice that were

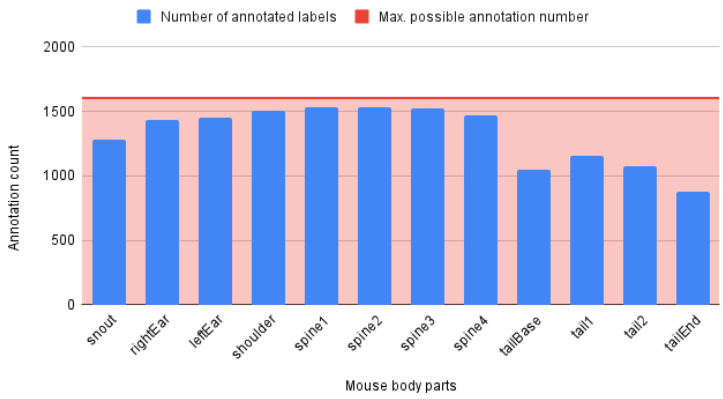


Figure 4.3: Total number of manually annotated labels for each body part compared to the maximum number of possible annotations across 535 different images, each containing 3 mice. 15,900 body parts were manually annotated using the DeepLabCut GUI. The *shoulder* and *spine* labels comprise the most often annotated body parts. while the *tailbase* and *tail* are annotated less frequently.

mostly actively moving. The evaluation was performed subjectively by visually inspecting the videos, frame by frame (Table **??**).

Property	Vid-10	VidFewIDSwaps	VidManyIDSwaps
Number of Videos	10	1	1
Individuals per Video	3	3	3
Total Duration (minutes)	1.0	3.2	2.0
Total Frames	15,000	4,785	3,000
Fur Color	Black / Gray / White	Black	Black
RFID Tagging	No	Yes	Yes

Table 4.1: Summary of the datasets used to evaluate model performance on previously unseen video recordings. The datasets differ in duration, number of frames, and identity-switching complexity, enabling assessment of tracking robustness and identity preservation under varying conditions.

4.3 Tracking performance evaluation

We evaluated the tracking performance on the Vid-10 dataset, which contains 10 1-minutes videos with three mice with different fur colors across video frames. Overall,

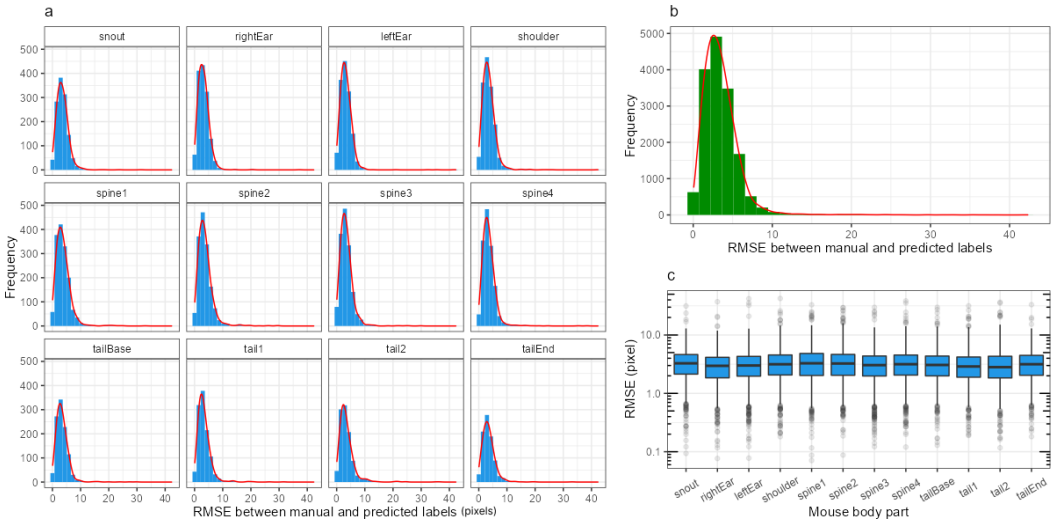


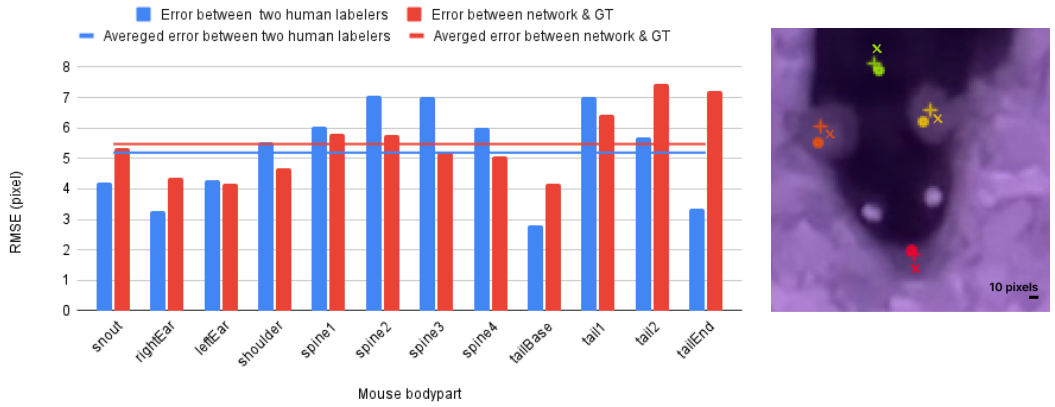
Figure 4.4: **a**, Histogram of the Root Mean Square Error (RMSE) between the network's predicted labels and their ground truth labels for each body part and a confidence likelihood of 0.9. **b**, RMSE for train and test data (all body parts included) with P-cutoff of 0.9. The error in **a** and **b** is normally distributed. Red lines (density) estimate the observed data. **c**, Boxplot of the distribution of the RMSE between manual and predicted labels for each body part (with P-cutoff of 0.9). The boxes span the range from the 25th to the 75th percentile. 75% of the network's predictions are below 4.47 pixels and 50% are less than 3.1 pixels. Black horizontal lines denote the median. Averaged median of all body parts is 3 pixels. Dots indicate outliers.

the averaged MOTA score of all videos was 0.98 which is consistent with the value reported by DeepLabCut. There were no false positive detections, 2.6% false negative detections, and 0.46% identity switches (Fig. 4.6). The Precision and recall values were 1.0 and 0.997, respectively.

4.4 Evaluation of mice identification using RFID

To evaluate the performance of the RFID-DLC matching method, we separately evaluated both the VidFewIDSwaps and VidManyIDSwaps datasets. Each dataset have one video with black mice, that are implemented with an RFID tag. We first assessed the videos on the trained model to obtain the keypoint predictions. In both videos, the network predicted 80% of all potential *snout* labels, 45% to 90% of all *tail* labels and above 90% of the remaining labels. (Fig. 4.7).

Then we applied the RFID-DLC matching method. Matching the centroid of the network's predicted mice with their corresponding RFID tag is based on the number of



(a) (b)

Figure 4.5: **a**, Blue bars indicate the RMSE between two human annotator's labels on 20 images (approximately 720 body parts) with three mice in each image (same images used in network test set). The *tailbase* label achieves a minimum error of 2.8 pixels, while the *spine* labels have an error between 6 and 7 pixels. The blue line indicates the average RMSE across all body parts (5.2 pixels). Red bars represent the RMSE between human labels and network's predicted labels on the same images, with a P-cutoff of 0.9 pixels. The network performance reached human level accuracy, reaching an error of 2.8 pixels on average (red line).**b**, An example of a high-magnification random chosen image, showing the snout, ears and shoulder labeled by a human experienced annotator (plus), inexperienced annotator (cross) and by the network (dot). Network prediction approximately matches the human manual annotations.

overlap in each frame, which depends on the length of the time interval for the RFID events (Fig. 3.4). In order to have an optimum overlap value and appropriate average location (coordinate in image space) of RFID events of each mouse caring an RFID tag within every time interval, we investigated several interval lengths (0.5, 1, 2, and 3 seconds) and calculated the average number of matches and mismatches, as well as the difference value between them.

In the VidFewIDSwaps dataset, for a one-second time interval, there was 44% RFID-DLC match, 4% mismatch, and the remaining 52% of the entire video resulted when either there were no RFID detections (FN) occurred or the mice were not present in the range if the RFID detections (FP) (Fig. 4.7a). For the VidFewIDSwaps dataset, the MOTA score of the RFID-DLC method was calculated to be 0.43. For the VidManyIDSwaps dataset there was 23% RFID-DLC matching, 8% mismatching (Fig. 4.7c) and the MOTA score for was 0.22. Fig. 4.7b and Fig. 4.7d show the total number of overlaps between all RFID tags and DLC dummy IDs pairs in a one-second time interval for the VidFewIDSwaps and VidManyIDSwaps datasets, respectively.

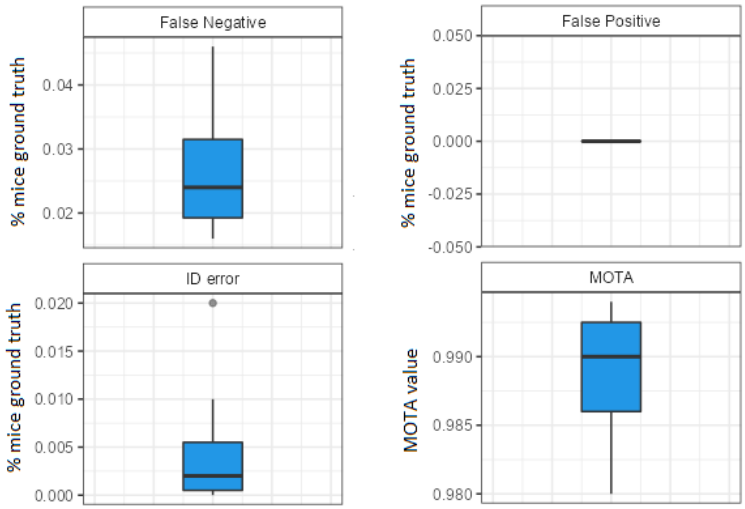
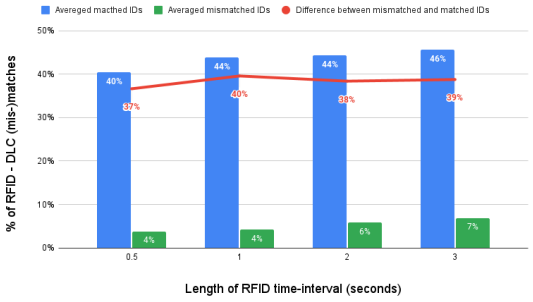


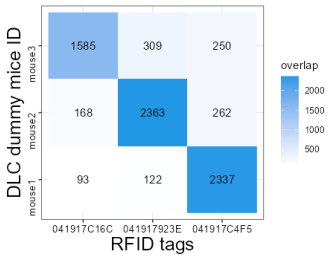
Figure 4.6: Percentage of false negative (FN), false positive (FP), and incorrectly matched detections (ID-Error) out of the total number of detections in 10 different random 1-minute (each 1500 frames) videos. Each data point corresponds to video. There were no FP detections in all videos. The average percentage of FN and ID-Error detections over all videos is 2.6% and 0.46%, respectively.

4.5 Performance of correcting ID-switches

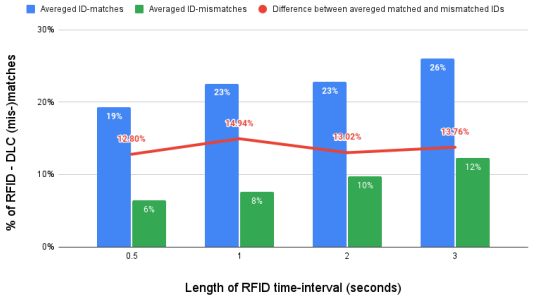
At this stage, the performance of CSI method is evaluated on the VidManyIDSwaps dataset. The MOTA score was calculated for the network's predicted labels, which then was 0.64. After performing the RFID-DLC matching procedure, we applied the CSI method and calculated the MOTA value again, which was increased by 0.01%. The number of mismatched IDs was reduced by 11% (-332 ID errors) while the total percentage of FN predictions increased from 1.7% to 5.4%. (Fig. 4.8).



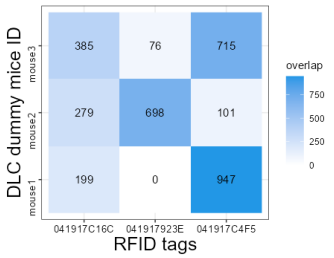
(a) Time intervals' statistic for VidFewIDSwaps dataset



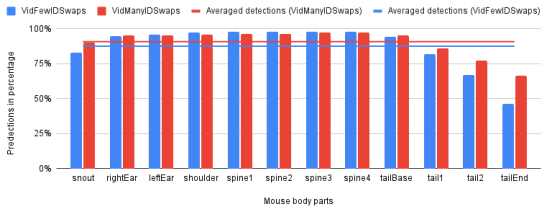
(b) Overlaps of all RFID-DLC detections in the VidFewIDSwaps dataset



(c) Time intervals' statistic for VidManyIDSwaps dataset



(d) Overlaps of all RFID-DLC detections in the VidManyIDSwaps dataset



(e)

Figure 4.7: **a, c**, Several time-interval lengths (in seconds). Blue bars represent the average percentage of matches, green bares indicate mismatches between the RFID tag detections and the centroid of the network's predicted labels, as well as the difference value between them (red curve). The interval with the length of one second achieved the maximum number of RFID-DLC matches and the minimum number of mismatches in both datasets: VidFewIDSwaps (**a**) and VidManyIDSwaps(**c**). **b, d**, Heatmaps showing the total number of overlaps for all RFID-DLC identity permutations for both datasets: VidFewIDSwaps (**b**) and VidManyIDSwaps(**d**). Color intensity indicates the probability (overlap) of an RFID-DLC matching. A predicted mouse by the network (with a dummy name) is more likely to be attached to an RFID tag when its overlap value with a specific tag is higher than the overlap value with the other RFID tags. **e**, Number of body parts (in percentage) were predicted by the network. Red and blue indicates the VidManyIDSwaps and VidFewIDSwaps, respectively. Lines represent their averaged value.

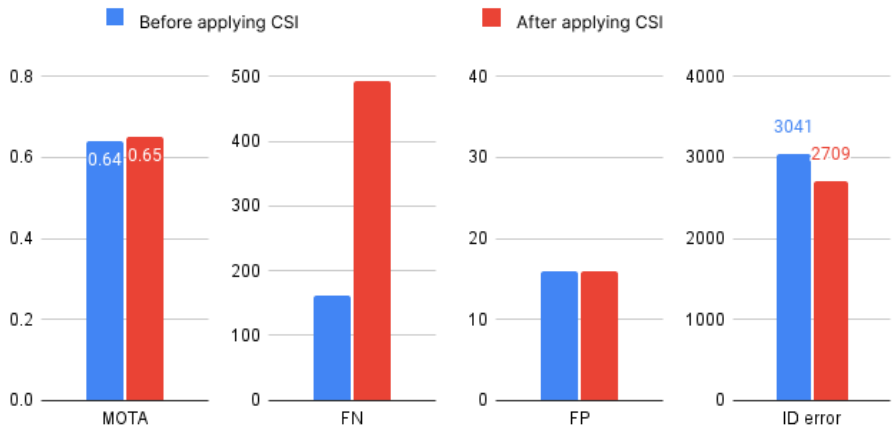


Figure 4.8: Tracking performance before and after utilizing the CSI method on the Vid-ManyIDSwaps dataset.

5 Discussion

In the visually inspected videos, we have seen that with few labeled images, DeepLabCut reached human-level accuracy. However, DeepLabCut shows issues with tracking multiple individuals in occluded situations. Therefore, besides DeepLabCut, we utilized an RFID (Radio Frequency Identification) system with mice that are attached with physical tagging. Such physical tagging is essential to identify and differentiate between various mice. RFID tagging can also be used as an automated method for periodic correction of mice identity switches [py] without the need for manual inspection. One limitation of employing RFID system for mice identification, is that RFID tag detections are not precise as the neural network predictions, especially when numerous mice are housed in an 820 cm^2 cage. Therefore, utilizing RFID system to correct identity switches may result in certain identity errors (ID_error), which occur when a mouse overlaps with a mismatched RFID tag, and FN detections, which can arise when just one of the switched Identities is corrected

We provide a set of scripts that combines DLC and the RFID system to perform pose estimation, tracking, and identification of a horde of interacting mice in a laboratory home cage. Researchers can use our software to record videos over long times scales and analyze novel videos from comparable experimental settings. We developed our scripts to be scalable, customizable and easy to use. In addition, we designed an automated method allowing the user to dynamically specify the RFID reader positions on the video frame with no prior-configuration. This is beneficial for ensuring consistency when changing the camera position. The RFID system we used consists of eight RFID readers; nevertheless, our scripts are capable to function with any number of RFID readers. Our trained model has been trained on images containing three mice, but this is not a strict upper bound. For inference, users can run the experiment with more or less mice.

5.1 Overall performance

The number of identity error can be decreased by adding more labeled images to the training set and resuming the training [ref]. Hence, we repeated the iterative training process, expanding our training dataset in each training round, until we finally got a better model which achieved good results on truly novel videos. The final model is feasible for research use without the need for further training data. Only 278 trained images (80% of entire labeled dataset) were required to reach less than 5.2 pixels test error

on an image size of 1280x720 pixel, whereas 428 trained images were able to achieve less than 4.9 pixels on test set, which nearly matches human validation accuracy (5.2 pixels). Furthermore, we measured the human variability in labeling the same images and claimed that the accuracy of labeling body parts between two human annotators is proportional to the accuracy of the network's predicted labels. Moreover, the averaged distance between two human annotators' labels and those predicted by the network is 0.2 pixels, which we consider to be an outstanding result (Fig. 4.5).

Because of occlusion (when two or more mice get very close and look combined with each other) some body parts have been less annotated (manually) than others (*tails* keypoints) (Fig 4.3), resulting in a training set with varied quantities of each keypoint. The prediction of a certain body part can perform poorer when the training set contains images with fewer annotations of this body part (Fig. 4.5), affecting the overall model performance.

Another factor that enhances the model is having consistent annotations in the training set. For some keypoints, it is not easy to define a convention for consistently annotating the same position. Keypoints such as *snout*, *ears* and *tailBase* are simple to annotate, as they have a consistent shape, and their skin tone color differs from the fur, whereas *spine* and *tail* are not.

During visually evaluating the model, we noticed that the network performed the worst on gray mice as they are hard to distinguish from the bedding (?), especially when they run fast in the dark phase, thus we annotated more images of this kind and merged them with the training set to use them in next training rounds (Table **??**).

We believe that the amount of images in the test set is insufficient to provide a reasonable judgement about the model performance (only 107 images). A feasible option is to evaluate the model on more annotated images, but this will be very time-consuming. However, model evaluation using the RMSE metric is insufficient for assessing the overall model performance, as it only calculates the distance error between train and test sets without considering missing predictions (FN) or switched identities (ID_error). Therefore, we utilized the MOTA and mAP metrics on new datasets.

All datasets established for the evaluation stage were visually inspected. The video evaluation revealed a high prediction accuracy in posture prediction, numerous predicted keypoints and a low value for FN predictions. The *ear*, *shoulder* and *spine* keypoints were the most accurately detected, whereas the *tail* labels were the most difficult. However, the videos were recorded in the dark phase (where mice are most active), which results in lower video quality and impacts the network predictions. The model performed very well on the Vid-10 dataset. The videos were not hard to analyze, as all detected keypoints were placed on the mice bodies (no FP detections), and very few missing detections (FN) and mismatched identities (ID_error). The VidManyIDSwaps data set was the most difficult to inspect accurately, as the mice identities often switched over long periods of time.

5.2 Identification of multiple mice

Our RFID-DLC matching method worked well on both Vid-Few-ID-Swaps and Vid-ManyIDSwaps datasets. We investigated three criteria that influence the ID-matching process is performed: 1) length of the time interval, 2) video length and mice activity and 3) number of ID switches. To find the optimal RFID tag to DLC dummy ID assignment, we select the time interval length (Fig 3.4.) that maximizes the number of overlaps of ID-matches and minimizes the overlap of ID-mismatches in all assignments. This corresponds to the greatest difference value between matches and mismatches of all examined interval lengths. We determined that the optimal interval length for both datasets is one second, as seen in Fig. 4.7a and 4.7c.However, experimenters can use the generated statistics to determine which interval length delivers the optimum performance. The video length in the VidFewIDSwaps dataset was sufficient to achieve a wide variation in the overlap value between ID-matches and ID-mismatches (Fig. 4.7b). Even though the ID-matching method has successfully assigned each DLC dummy ID with its respective RFID tag (which must be unique) in the VidManyIDSwaps dataset, the numbers of ID-matches and ID-mismatches for each RFID-DLC pair is very close. The video in the VidManyIDSwaps dataset has several switched identities that propagate throughout the rest of the video, resulting in an increase in the overlap value for mismatched identities (Fig. 4.7d). If two or more DLC identities have their highest overlap with the same RFID tag, assigning a unique RFID tag to each DLC mouse will be permitted, and the user will be prompted to supply a longer or shorter video. However, the number of mice in the video can affect the performance of the ID-matching method, as the algorithm increases the overlap value for each DLC mouse presented in the same RFID region of a detected RFID tag. Thus, the ID-matching algorithm is more confident when fewer mice presents in the video. Another consideration should be taken into account when applying the ID-matching method is to ensure having a low number of FN detections (unlabeled mice), as this increases the overlap with matched IDs. The number of FN detections is proportional to the DLC P-cutoff value.

5.3 Correction of switched identities: solution and limitations

The CSI method has detected 343 potential identities switches, of which 332 (96%) resulted in correction of the assigned IDs and 4% produced incorrect ID assignment (there were no switches). An incorrect ID reassignment (false correction) occurs as a result of imprecise RFID detections; when a DLC mouse is present in the same range of an RFID tag that is matched with another DLC mouse. When detecting potential identity switches, CSI is not swapping IDs but rather reassigning IDs, to not create any

false ID assignment (because it is not possible to know between which two individuals to swap the IDs) but resulting in more FN detections (Fig 4.8) which is the reason why the MOTA score does not decrease significantly after using the method.

6 Conclusion

We utilized the DeepLabCut toolbox to track and localize several mice keypoints in complex social interactions without the need for physical markers.

Besides DeepLabCut, we utilized an RFID (Radio Frequency Identification) system to identify and differentiate between various mice, that are microchipped with an RFID tag. Furthermore, we used RFID tagging to automate periodic correction of mice identity switches without the need for manual inspection. One limitation of employing RFID system in order to identify multiple mice, is that the RFID tag detections are not precise as the neural network predictions, especially when numerous mice are housed in an 820 cm^2 cage. Therefore, utilizing an RFID system to correct identity switches may result in certain identity errors, which occur when a mouse overlaps with a mismatched RFID tag, and FN detections, which can arise when just one of the switched Identities is corrected.

In our experiment, we provide a set of scripts that combines DLC and the RFID system to perform pose estimation, tracking, and identification of a horde of interacting mice in a laboratory home cage. Researchers can use our software to record videos over long times scales and analyze novel videos from comparable experimental settings. We developed our scripts to be scalable, customizable and easy to use. In addition, we designed an automated method allowing the user to dynamically specify the RFID reader positions on the video frame with no prior-configuration. This is beneficial for ensuring consistency when changing the camera position. The RFID system we used consists of eight RFID readers; nevertheless, our scripts are capable to function with any number of RFID readers. Our trained model has been trained on images containing three mice, but this is not a strict upper bound. For inference, users can run the experiment with more or less mice.

Bibliography

[1] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. 2017. Understanding of a convolutional neural network. In *2017 international conference on engineering and technology (ICET)*. Ieee, 1–6.

[2] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 8 (2013), 1798–1828.

[3] Keni Bernardin and Rainer Stiefelhagen. 2008. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing* 2008 (2008), 1–10.

[4] Fabrice de Chaumont, Renata Dos-Santos Coura, Pierre Serreau, Arnaud Cressant, Jonathan Chabout, Sylvie Granon, and Jean-Christophe Olivo-Marin. 2012. Computerized video analysis of social interactions in mice. *Nature methods* 9, 4 (2012), 410–417. https://doi.org/10.1038/nmeth.1924doi:10.1038/nmeth.1924

[5] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.

[6] Jiawei Han, Micheline Kamber, and Jian Pei. 2012. Data Mining Trends and Research Frontiers. In *Data Mining*. Elsevier, 585–631. https://doi.org/10.1016/B978-0-12-381479-1.00013-7doi:10.1016/B978-0-12-381479-1.00013-7

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. https://doi.org/10.48550/arXiv.1512.03385doi:10.48550/arXiv.1512.03385

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[9] Pierre Karashchuk, Katie L. Rupp, Evyn S. Dickinson, Sarah Walling-Bell, Elischa Sanders, Eiman Azim, Bingni W. Brunton,

and John C. Tuthill. 2021. Anipose: A toolkit for robust markerless 3D pose estimation. *Cell reports* 36, 13 (2021), 109730. https://doi.org/10.1016/j.celrep.2021.109730doi:10.1016/j.celrep.2021.109730

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[11] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep Learning. *Nature* 521, 7553 (2015), 436–444.

[12] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

[13] Tanmay Nath, Alexander Mathis, An Chi Chen, Amir Patel, Matthias Bethge, and Mackenzie Weygandt Mathis. 2019. Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nature protocols* 14, 7 (2019), 2152–2176. https://doi.org/10.1038/s41596-019-0176-0doi:10.1038/s41596-019-0176-0

[14] M. Pistacchi, M. Gioulis, F. Sanson, E. de Giovannini, G. Filippi, F. Rossetto, and S. Zambito Marsala. 2017. Gait analysis and clinical correlations in early Parkinson's disease. *Functional neurology* 32, 1 (2017), 28–34. https://doi.org/10.11138/FNeur/2017.32.1.028doi:10.11138/FNeur/2017.32.1.028

[15] Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big Data* 3, 1 (2016). https://doi.org/10.1186/s40537-016-0043-6doi:10.1186/s40537-016-0043-6