vmTracking: Virtual Markers Overcome Occlusion and Crowding in Multi-Animal Pose Tracking

Hirotsugu Azechi[1*], Susumu Takahashi[1]

[1]Laboratory of Cognitive and Behavioral Neuroscience, Graduate School of Brain Science, Doshisha University, Kyotanabe, 610-0394, Japan.

*Correspondence: hazechi@mail.doshisha.ac.jp

Abstract

Overcoming occlusion and crowding in multi-animal tracking remains challenging. Thus, we aim to introduce virtual marker tracking (vmTracking) as a solution to these problems. This method integrates markerless multi-animal pose estimation, including multi-animal DeepLabCut (maDLC) or Social LEAP Estimate Animal Poses (SLEAP), with single-animal tracking techniques, such as DeepLabCut and LEAP. Initially, maDLC or SLEAP is employed to create videos in which the tracking results are labeled as "virtual markers." Subsequently, these virtual markers are used to ensure consistent animal identification across frames, and the animals are tracked using single-animal tracking. vmTracking significantly reduces the need for manual corrections and the number of annotation frames required for training. It more efficiently improves occlusion and crowding compared with traditional markerless multi-animal tracking, which relies on motion state estimation. The findings of this research have the potential to substantially enhance the precision and dependability of tracking methods used in the analysis of intricate naturalistic and social behaviors in animals.

Keywords: multi-animal tracking, Virtual marker, occlusion, crowding, DeepLabCut, SLEAP

Introduction

Deciphering the complex dynamics of animal behavior is a cornerstone to advancing our understanding of biological and neurological processes. Central to this quest is the study of naturalistic and social animal behavior, which demands precise tracking and pose estimation of multiple individuals simultaneously. Advancements in machine learning have brought about sophisticated tools that can be used to analyze the behavior of single animals with remarkable accuracy. These tools include DeepLabCut (DLC; Mathis, et al., 2018) and LEAP, which estimates animal pose (LEAP; Pereira et al., 2019). Subsequently, the multi-animal version of DLC (maDLC; Lauer, et al., 2022) has been developed to meet the growing needs of social behavior research, allowing for the pose estimation of several individuals within a group. In parallel, the Social LEAP (SLEAP; Pereira et al., 2022) framework has emerged, presenting an independent approach with versatile neural network architectures designed to be species-agnostic and adaptable to various behavioral patterns. Current multi-animal tracking systems incorporate movement prediction strategies to maintain consistent identification of individuals across frames, utilizing techniques such as Kalman filters to predict and track animal movements (Yin et al., 2020).

However, the leap to multi-animal tracking has led to new challenges, such as occlusion and crowding. Occlusion occurs when one animal obscures another from the camera's view. Crowding is when the density of animals in a confined space is high, leading to errors in identifying and tracking specific animals. Despite efforts in research and development to resolve occlusion and crowding (Dendorfer et al., 2020; Han et al., 2023), these issues remain unresolved and hinder accurate multi-animal tracking and pose estimation.

Although the implementation of movement-based predictive tracking has enabled the tracking of multiple animals, occlusion and crowding continue to impede the seamless tracking of individuals, often necessitating manual intervention to ensure data integrity. This highlights the need for technological improvements in this field. Existing strategies are effective in many scenarios, but they still struggle to cope efficiently with occlusion and crowding, particularly in the analysis of extensive datasets. Therefore, researchers must exercise caution and frequently engage in manual verification to validate the tracking data. This reality signifies an ongoing demand for technological advancements in multi-animal tracking and pose estimation. Thus, this study aimed to explore current methodologies, evaluate the performance of existing systems, and suggest avenues for improvement. The results of this study could revolutionize the accuracy and reliability of tracking in analyzing complex naturalistic and social animal behavior.

Results

To refine multi-animal tracking techniques, we developed an innovative virtual marker method called "vmTracking." This approach entails attaching virtual markers to markerless animals to accurately track and identify them (Figure 1). Our initial step involved conducting comprehensive markerless multi-animal tracking and subsequently producing a video that captured the tracking results. We designated these output labels as "virtual markers," differentiating them from conventional physical markers. By analyzing these virtual marker videos, we meticulously annotated the identifications (IDs), leveraging the virtual markers for precise ID. Furthermore, we applied the method previously used for single-animal tracking to this technology, thereby extending its scope to include multi-animal tracking. We rigorously applied this methodology to DLC and SLEAP platforms and performed a comparative analysis to evaluate the accuracy of our vmTracking, which includes single-animal DLC (saDLC) (VMT-DLC) and LEAP (VMT-LEAP), against established multi-animal tracking methods (maDLC or SLEAP).

Enhanced tracking of black mice against black backgrounds: A virtual marker approach

To accurately track black C57BL/6J mice against a black background, we focused on refining our tracking techniques to enhance contrast and ID accuracy. We conducted an experiment to track five mice with six keypoints per mouse. Markerless multi-animal tracking was performed using maDLC or SLEAP (Figure. 2a and b). Subsequently, we created a video using SLEAP, outputting two keypoints as virtual markers (Figure 2a and c), and conducted VMT-DLC and VMT-LEAP (Figure 2d). The results stem from five iterations of retraining for both maDLC and VMT-DLC, with the final total number of annotated frames being 1097 and 1196, respectively. The outcomes for SLEAP and VMT-LEAP are based on training and predictions by the import of annotation data from maDLC and VMT-DLC. Our analysis of the four tracking methods revealed significant improvements in all five evaluation metrics (Figure S1) (maDLC vs. VMT-DLC: Matches: $p < 0.001$, false negatives (FN): $p < 0.001$, false positives (FP): $p < 0.05$, ID switches: $p < 0.001$, root mean square error (RMSE): $p < 0.05$, Greenhouse–Geisser-corrected, repeated-measures analysis of variance (ANOVA), $n = 12$, Figure 2e). Specifically, VMT-DLC significantly enhanced the accuracy of matching keypoints (Matches: $p < 0.01$) and reduced FN ($p < 0.001$), but increased FP ($p < 0.05$). No significant differences were found for ID switches (n.s) and RMSE (n.s) (Figure 2e). When comparing SLEAP and VMT-LEAP, significant improvements were noted not only in Matches and FN but also in FP and RMSE (SLEAP vs VMT-LEAP: Matches: $p < 0.01$, FN: $p < 0.01$, FP: $p < 0.001$, ID switches: n.s, RMSE: $p < 0.001$, Figure 2e). However, in the comparison of vmTracking, VMT-DLC showed higher accuracy in Matches, FN, and ID switches (VMT-DLC vs VMT-LEAP: Matches: $p < 0.001$, FN: $p < 0.001$, FP: n.s, ID switches: $p < 0.001$, RMSE: n.s, Figure 2e).

In this study, the criterion for a Match is stringent, requiring a match in both ID and body part.

Consequently, even if tracking after an ID switch is accurate, it is not considered a Match. Therefore, an evaluation was conducted by further classifying into FN where the correct body part was predicted but the ID was incorrect (ID mismatch but Bodypart match FN), and FN where the correct body part could not be predicted, including missing predictions (Bodypart mismatch FN) (see Figure S3). Repeated measures ANOVA showed significant changes in both ID mismatch but Bodypart match FN ($p < 0.01$) and Bodypart mismatch FN ($p < 0.001$) (Greenhouse–Geisser-corrected, repeated-measures ANOVA, $n = 12$, Figure S3). In pairwise tests, no significant differences were observed in ID mismatch but Bodypart match FN (n.s). However, vmTracking generally maintained these at a low level (Figure S3a). In contrast, vmTracking significantly reduced Bodypart mismatch FN (maDLC vs VMT-DLC: $p < 0.001$, SLEAP vs VMT-LEAP: $p < 0.01$), with VMT-DLC showing a significant reduction compared to VMT-LEAP ($p < 0.001$, Figure S3b).

The correlation between the ID match of virtual markers (Figure S2) and the matching rates of vmTracking was depicted using a scatter plot (Figure 2f), where both VMT-DLC ($r = 0.660$, $p < 0.05$) and VMT-LEAP ($r = 0.853$, $p < 0.001$) were positively correlated with virtual marker accuracy. Notably, VMT-LEAP achieved a matching rate (77.4%) comparable to virtual marker accuracy (73.9%), whereas VMT-DLC exceeded it with a matching rate of 91.4%. Therefore, vmTracking enhanced multi-animal tracking accuracy, particularly using DLC, in challenging contrast scenarios (Supplementary Video 1).

Refining tracking accuracy in varied conditions

To validate the data of vmTracking under diverse conditions, we conducted an experiment with three mice against a white background, focusing on scenarios classified as occlusion-congested (OC) or non-occlusion-congested (nOC). This differentiation allowed a comprehensive assessment of vmTracking at varying levels of environmental complexity. Following our established protocol, we assigned six keypoints per mouse for markerless tracking (Figure 3a and b, left; Figure 4a and b, left). We then generated videos using maDLC that displayed two keypoints as virtual markers (Figure 3a, Figure 4a), followed by vmTracking (Figure 3b right, Figure 4b right). The outcomes for maDLC and VMT-DLC were obtained after 11 iterations of retraining (with a final total of 3886 annotated frames) and 5 iterations of retraining (with a final total of 1167 annotated frames), respectively. In both OC and nOC scenes, vmTracking demonstrated significant improvements in match accuracy (OC: $p < 0.001$, nOC: $p < 0.05$, Wilcoxon signed-rank test, OC: $n = 12$, nOC: $n = 8$) and FN (OC: $p < 0.001$, nOC: $p < 0.05$), while maintaining steady results in FP, ID switches, and RMSE (OC: n.s, nOC: n.s) (Figure 3c and 4c). Remarkably, vmTracking not only enhanced the tracking accuracy in OC scenes but also improved it in nOC scenes, where maDLC already showed a high matching rate (Figure 4c Matches). Additionally, in these conditions, a reduction in ID mismatch but Bodypart match FN was observed in OC scenes (OC: $p < 0.01$, nOC: n.s, Figure S4a

and S5a), and Bodypart mismatch FN significantly decreased with vmTracking (OC: $p < 0.001$, nOC: $p < 0.05$, Figure S4b and S5b).

The correlation between the accuracy of the virtual markers and the matching rate of VMT-DLC was depicted using a scatter plot (Figure S6). This analysis revealed that when combining OC scenes ($n = 12$) and nOC scenes ($n = 8$), there was no significant positive correlation between the accuracy of virtual markers and the matching rate ($r = 0.434$, n.s, $n = 20$ in total). However, in OC scenes, VMT-DLC achieved a matching rate of 95.2%, surpassing the virtual marker accuracy of 89.5%. In nOC scenes, the matching rate was 97.5%, which is comparable to the virtual marker accuracy of 98.4%. Thus, VMT-DLC effectively enhanced the accuracy of multi-animal tracking in challenging scenarios, such as scenes with occlusion and crowding (Supplementary Video 2 and 3).

Comparison of the number of annotation frames between VMT-DLC and maDLC

On average, VMT-DLC achieved an accuracy rate of 95.2% ($n = 12$) in tracking three mice in OC scenes with only 1,167 annotation frames (Figure 5), significantly outperforming maDLC, which required approximately 13,732 annotation frames for comparable accuracy. This result underscored the efficiency of VMT-DLC in providing precise tracking with fewer training samples.

Discussion

In this study, we introduce vmTracking, a novel approach that effectively addresses occlusion and crowding in multi-animal tracking. Our experimental results demonstrate that vmTracking significantly enhances accuracy compared with traditional tracking methods, especially in complex scenarios. In this section, we discuss the advantages and limitations of vmTracking, as well as its implications for future research directions and on this field.

Despite advancements in markerless multi-animal tracking and pose estimation technologies, occlusion and crowding remain challenging to overcome. Further advancements are essential to address these persistent issues effectively (Lauer et al., 2022). One potential solution involves attaching identifiable physical markers to animals as tracking cues. However, physical markers have potential impact on animal behavior. Thus, technologies that solve this challenge without relying on physical markers need to be developed. In the present study, we successfully improved the accuracy of traditional multi-animal pose estimation, even in scenarios involving occlusion and crowded scenes, by adopting an extremely simple strategy of applying virtual markers to videos of markerless animals, thereby eliminating the need for physical markers.

Whether based on DLC or SLEAP, vmTracking performed better than traditional markerless multi-animal tracking methods. This improvement in accuracy was observed in almost all scenes, regardless of the experimental conditions (tracking three or five animals) or scenarios (occlusion and

crowding). In particular, with VMT-DLC, scenes that previously had a keypoint matching rate of less than 50% in maDLC achieved approximately 90%, indicating a remarkable improvement in accuracy. Although the performance of VMT-SLEAP was inferior to that of VMT-DLC, it was still clearly superior to that of SLEAP. These results indicate that vmTracking is a robust solution to occlusion and crowding.

The accuracy of vmTracking was generally positively correlated with the accuracy of the virtual markers, suggesting that vmTracking relies on virtual markers. Therefore, investing effort into improving the accuracy of virtual markers can increase the precision of outcomes. However, vmTracking might be effective but not necessarily an efficient method if the accuracy cannot be guaranteed without sufficient labeling of the virtual markers on the tracked individuals. Yet, in reality, extensive labeling is not necessary. In the case of VMT-DLC, the ID matched virtual markers was between 70% and 80%, with the final accuracy of vmTracking exceeding 90%. Importantly, the accuracy of the virtual markers in the present study was based on whether a label was attached to any part of an individual with a matching ID, which was a less stringent standard than matching keypoints in tracking results. For the creation of virtual markers, it is sufficient to ensure that the IDs of the keypoints to be marked are correct and that some errors are permissible. Although accurate results in markerless multi-animal tracking require addressing keypoints in all frames, the effort required for corrections in creating virtual markers is considerably less.

Irrespective of the tracking methodology employed, increasing the number of annotation frames is expected to enhance performance. However, achieving commendable results with fewer training frames is desirable. vmTracking also demonstrates exceptional performance in terms of the number of frames required for training. In experiments involving the tracking of three mice, VMT-DLC achieved a keypoint matching rate of over 90%, with approximately 1200 annotation frames, even in scenes with occlusion and crowding. By contrast, maDLC would need to annotate more than 10 times the number of frames to attain similar accuracy, an undertaking that is impractical. Furthermore, this estimation is based on simple regression, and in reality, the accuracy may become saturated. Thus, VMT-DLC can achieve with minimal training what is extremely difficult to attain using maDLC.

In conclusion, the development and implementation of vmTracking in this study marks a significant advancement in markerless multi-animal tracking. The superior accuracy and efficiency of vmTracking, particularly in dealing with occlusion and crowding, demonstrate its potential as a robust alternative to traditional tracking methods. It is important to acknowledge, however, that in scenarios where the tracked individuals are entirely obscured, maintaining this high level of accuracy may not be guaranteed. The reliance of vmTracking on virtual markers, which significantly reduces the need for extensive manual annotation and training, further underscores its practicality and potential for widespread application. Moreover, the ability of vmTracking to achieve high accuracy

with a relatively low number of annotation frames is a substantial improvement over existing methods. The findings of this study not only contribute to the refinement of animal tracking technologies but also open avenues for future research, particularly in enhancing the precision of virtual markers and extending the application of vmTracking to diverse and complex environments. Thus, vmTracking has emerged as a transformative tool in the study of animal behavior, ecology, and other related fields, offering a method that is effective and efficient in addressing some of the most persistent challenges in multi-animal tracking.

Methods

Multi-mouse dataset

Behavioral data were collected from C57BL/6J mice. These mice were grouped in cages of 2–3 each, in an environment with a temperature of 24–26°C and a 12-hour light-dark cycle. They had unrestricted access to food and water. All behavioral recordings were performed during the light phase. Cages and housing conditions were refreshed weekly. All procedures were approved by the Institutional Animal Care and Use Committee of Doshisha University.

Data were collected in two environments. In one setting, five C57BL/6J mice were allowed to roam freely in a rectangular open field with a black floor measuring 55 × 60 cm and a height of 40 cm. The recordings were conducted for approximately 30 min at a resolution of 640 × 640 pixels and a frame rate of 30 fps. In another setting, three C57BL/6J mice were allowed to roam freely in a circular open field with a white floor (31 cm in diameter) and a transparent wall (17 cm in height). The recording in this environment was performed for approximately 60 min at a resolution of 460 × 460 pixels and a frame rate of 30 fps.

DLC as markerless multi-animal tracking

In this study, DeepLabCut version 2.2.3 was employed. maDLC project was conducted to track and estimate the postures in the interaction behavior videos of either five or three mice. For annotation, 50 frames were extracted, and six keypoints (along the midline from the nose to the base of the tail) were labeled. Skeletons were created for all keypoint pairs. The training network used was DLCRnet, which is known for its high performance in tracking multiple animals (Lauer et al., 2022). After training, performance evaluation and video analysis were conducted. Tracking videos showing the predicted results were created, and the results were reviewed. Based on the results, outlier frames were extracted and annotated, and further training was performed. In this study, these refinement and training processes were repeated five times for the experiments involving tracking five mice, resulting in annotations for a total of 1097 frames. For the experiments tracking three mice, these processes were repeated eleven times, culminating in annotations for a total of 3886

frames. Coordinate data were obtained after each training and utilized to assess the accuracy of the tracking.

Creating virtual marker video using DLC

Virtual marker videos for vmTracking of three mice were created based on the results from iteration-6 of maDLC, which involved the annotation of 2820 frames. maDLC generates fragments of motion trajectories (tracklets) based on motion estimation. These tracklets were stitched together to form continuous behavioral trajectories for each animal. However, tracklet generation can be interrupted by occlusion, and stitching may fail, resulting in ID switches. Such ID switches were manually corrected. Corrections were also made for partial ID switches at the six keypoints for each mouse. In cases where there was a temporary switch in ID or when the predicted location of the keypoints was off but the predicted ID was correct, corrections were not always made. These corrected data were used to create tracking videos, which were then designated as virtual marker videos. These videos were created in grayscale, with the marker size set to 2, and only keypoints 2 and 4 of each mouse were output without skeletons.

vmTracking using DLC

saDLC project was created to analyze virtual marker videos of five or three mice. The number of keypoints per animal was set to six, and the total number of keypoints to be labeled was determined by multiplying this number by the number of tracked mice. Six keypoints were allocated per animal, and skeletons were constructed for all keypoint pairs for each animal. Similar to maDLC, 50 frames were extracted for annotation. For the training network, EfficientNet_b0 was selected, which achieved superior performance and lower computational costs than ResNet or MobileNet (Tan and Le, 2019). After training, a performance evaluation and video analysis were conducted, and tracking videos showing the predicted results were created and reviewed. Based on the results, outlier frames were extracted and annotated, and further training was performed. In this study, these refinement and training processes were repeated five times for the experiments involving tracking five mice, resulting in annotations for a total of 1196 frames. For the experiments tracking three mice, these processes were repeated five times, culminating in annotations for a total of 1167 frames. Finally, the training results were output to the markerless original videos to obtain tracking videos with the virtual markers removed.

Social LEAP as markerless multi-animal tracking

In this study, SLEAP version 1.3.3 was employed. The annotation data used in maDLC were imported, and training and prediction were conducted using SLEAP. The training type employed a “multi-animal top-down” approach, and maximum instances were set to match the number of

animals being analyzed. In prediction, a simple tracker method was selected, and the maximum number of tracks was set equal to the number of animals analyzed. The obtained coordinate data were used to evaluate tracking accuracy. Furthermore, tracking videos were created using the prediction results.

Creating virtual marker video using SLEAP

Virtual markers for the five-animal tracking experiment were created based on the SLEAP results. From the predicted outcomes of markerless multi-animal tracking, all predicted keypoints, except for keypoints 2 and 4, which served as virtual markers, were deleted. Subsequently, corrections were made to the ID switches and keypoint positions. The frames in which only one mouse was predicted were extracted. Annotations were added to the four unpredicted mice, followed by retraining and reprediction. Furthermore, frames in which only two mice were predicted were extracted. Random frames were selected, annotations for three unpredicted mice were added, and additional training and predictions were performed. In instances of temporary ID switches or when the predicted ID was correct despite a misalignment in the keypoint prediction position, they were not necessarily corrected. At this stage, labels for the unpredicted mice were not added. Using these corrected data, tracking videos were created and designated as virtual marker videos. During video creation, the marker size was set to two, and the skeletons were not output.

vmTracking using LEAP

The annotation data used in vmTracking with saDLC (VMT-DLC) were imported for training and prediction. The training type employed the “single animal,” that is, equivalent of LEAP. For prediction, a simple tracker method was selected, and the maximum number of tracks was limited to one. The obtained coordinate data were utilized to assess tracking accuracy. In addition, tracking videos displaying the predicted results were created.

Tracking performance evaluation

The accuracy of multi-animal tracking was evaluated based on a manually created ground truth for sampling sets of either 150 or 300 frames. This evaluation focused on tracking keypoints. To compare predicted data with ground truth, each predicted mouse ID was paired with the most matching ground truth ID. For this purpose, distances for all combinations of predicted data and ground truth IDs were calculated, and the Hungarian algorithm was used to determine the most suitable combinations. The evaluation metrics included Match, FN, FP, ID switch, and RMSE. Match, synonymous to the percentage of correct keypoints, is defined as the number of instances in which the predicted point is detected within a predefined threshold (10 pixels) from the ground truth point (Figure S1). This requires a match in both body part and ID with the ground truth. Cases not

qualifying as Match were assessed as FN. FN represents keypoints where correct predictions were not made, defined as either predictions within the threshold for an uncorresponding ground truth point or missing points (Figure S1). Furthermore, regarding FN, they were classified into two categories: one where the ID differed but the predicted part matched (ID mismatch but Bodypart match FN), and another where the predicted part was incorrect irrespective of the ID (Bodypart mismatch FN). FP represents keypoints that predict non-existent points, defined as points detected outside the threshold from all ground truth points (Figure S1). An ID switch is defined as an instance in which predictions are allocated within the threshold to any ground truth point of a different mouse across frames (Figure S1). RMSE was calculated based on pairs of associated predicted data and ground truth data.

Accuracy of virtual markers

The accuracy of the virtual markers was defined based on the identity matching rate for keypoints 2 and 4 (points becoming virtual markers) after correcting markerless multi-animal tracking results. In other words, an ID match was considered when the indicated ID of the virtual marker matched any ground truth for the same mouse within a threshold (10 pixels) (Figure S2). The accuracy of the virtual markers calculated as the ratio of virtual markers with matching IDs. If keypoints designated to become virtual markers were missing, the corresponding virtual markers will also be absent. These missing virtual markers were treated as mismatches.

Statistical analysis

In the tracking experiment with five mice, 12 OC scenes, in which the keypoints of different mice intersected, were sampled from four independent videos. The tracking accuracy based on the aforementioned metrics for maDLC, VMT-DLC, SLEAP, and VMT-LEAP was assessed through repeated-measures ANOVA and multiple comparisons with Bonferroni correction. In the tracking experiment with three mice, 12 OC and eight nOC scenes were sampled from four independent videos. The tracking accuracies of maDLC and VMT-DLC were compared using the Wilcoxon signed-rank test. In addition, a correlation analysis was conducted between the matching rate of vmTracking and the accuracy of the virtual markers based on the ID matching rate. Furthermore, for the OC scenes in the three-mouse tracking experiment, a regression analysis using the least-squares method was conducted to estimate the number of annotation frames required for maDLC to reach the accuracy level of VMT-DLC based on the matched keypoint rate. MATLAB functions ranova, multcompare, signrank, corrcoef, and polyfit were used in these analyses. Statistical significance was considered at $p < 0.05$

References

1. Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., & Bethge, M. (2018). DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning. Nature Neuroscience, 21(9), 1281–1289. https://doi.org/10.1038/s41593-018-0209-y

2. Pereira, T. D., Aldarondo, D. E., Willmore, L., Kislin, M., Wang, S. S.-H., Murthy, M., & Shaevitz, J. W. (2019). Fast animal pose estimation using deep neural networks. Nature Methods, 16(1), 117–125. https://doi.org/10.1038/s41592-018-0234-5

3. Lauer, J., Zhou, M., Ye, S., Menegas, W., Schneider, S., Nath, T., Rahman, M. M., Di Santo, V., Soberanes, D., Feng, G., Murthy, V. N., Lauder, G., Dulac, C., Mathis, M. W., & Mathis, A. (2022). Multi-animal pose estimation, identification and tracking with DeepLabCut. Nature Methods, 19(4), 496–504. https://doi.org/10.1038/s41592-022-01443-0

4. Pereira, T. D., Tabris, N., Matsliah, A., Turner, D. M., Li, J., Ravindranath, S., Papadoyannis, E. S., Normand, E., Deutsch, D. S., Wang, Z. Y., McKenzie-Smith, G. C., Mitelut, C. C., Castro, M. D., D'Uva, J., Kislin, M., Sanes, D. H., Kocher, S. D., Wang, S. S.-H., Falkner, A. L., … Murthy, M. (2022). SLEAP: A deep learning system for multi-animal pose tracking. Nature Methods, 19(4), 486–495. https://doi.org/10.1038/s41592-022-01426-1

5. Yin, X., Wang, J., & Song, S. (2021). Multi-target Tracking Based on YOLOv3 and Kalman Filter. In Q. Liang, W. Wang, X. Liu, Z. Na, X. Li, & B. Zhang (Eds.), Communications, Signal Processing, and Systems (Vol. 654, pp. 905–911). Springer Singapore. https://doi.org/10.1007/978-981-15-8411-4_121

6. Dendorfer, P., Ošep, A., Milan, A., Schindler, K., Cremers, D., Reid, I., Roth, S., & Leal-Taixé, L. (2020). MOTChallenge: A Benchmark for Single-Camera Multiple Target Tracking (arXiv:2010.07548). arXiv. http://arxiv.org/abs/2010.07548

7. Han, S., Wang, H., Yu, E., & Hu, Z. (2023). ORT: Occlusion-robust for multi-object tracking. Fundamental Research, S2667325823000328. https://doi.org/10.1016/j.fmre.2023.02.003

8. Tan, M., & Le, Q. V. (2020). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks (arXiv:1905.11946). arXiv. http://arxiv.org/abs/1905.11946

Acknowledgements

This work was supported by JSPS KAKENHI (21H05296 to S. T., 21H04247 to H. A.) and JST CREST (JPMJCR23P2 to S. T.).

Author contributions

H. A. and S. T. conceived the study. H. A. acquired the behavioral data of the mice. H. A. performed the animal tracking. H. A. performed the formal analyses. H. A. and S. T. prepared the manuscript.

Declaration of interests

The authors declare no competing interests.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this study, the authors used ChatGPT to improve readability and language. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

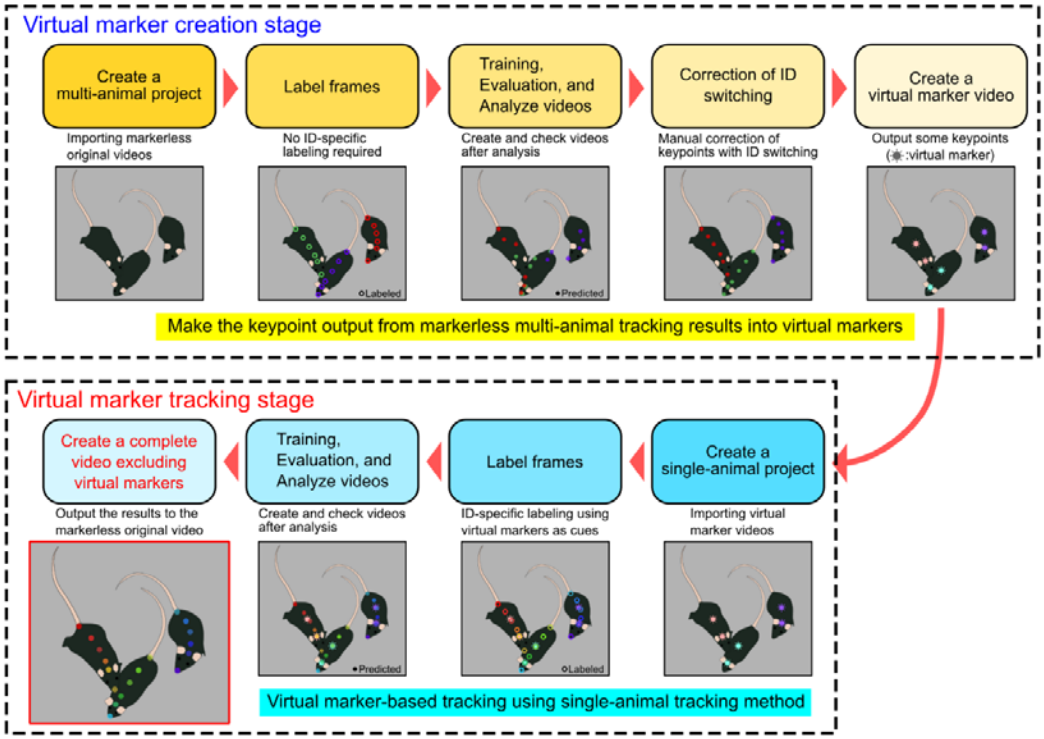

Figure 1. vmTracking workflow.

vmTracking comprises two stages: creating virtual marker videos and tracking the virtual markers. Traditional markerless multi-animal tracking was conducted during the video creation step. The results were manually corrected for ID switches and keypoints, after which a video displaying the selected keypoints was produced. These keypoints served as virtual markers for the subsequent tracking phases. During the tracking stage, these virtual markers were used as identifiers for annotation using single-animal tracking techniques. Finally, the tracking results were overlaid onto the original markerless video, thus generating a multi-animal tracking video without virtual markers.

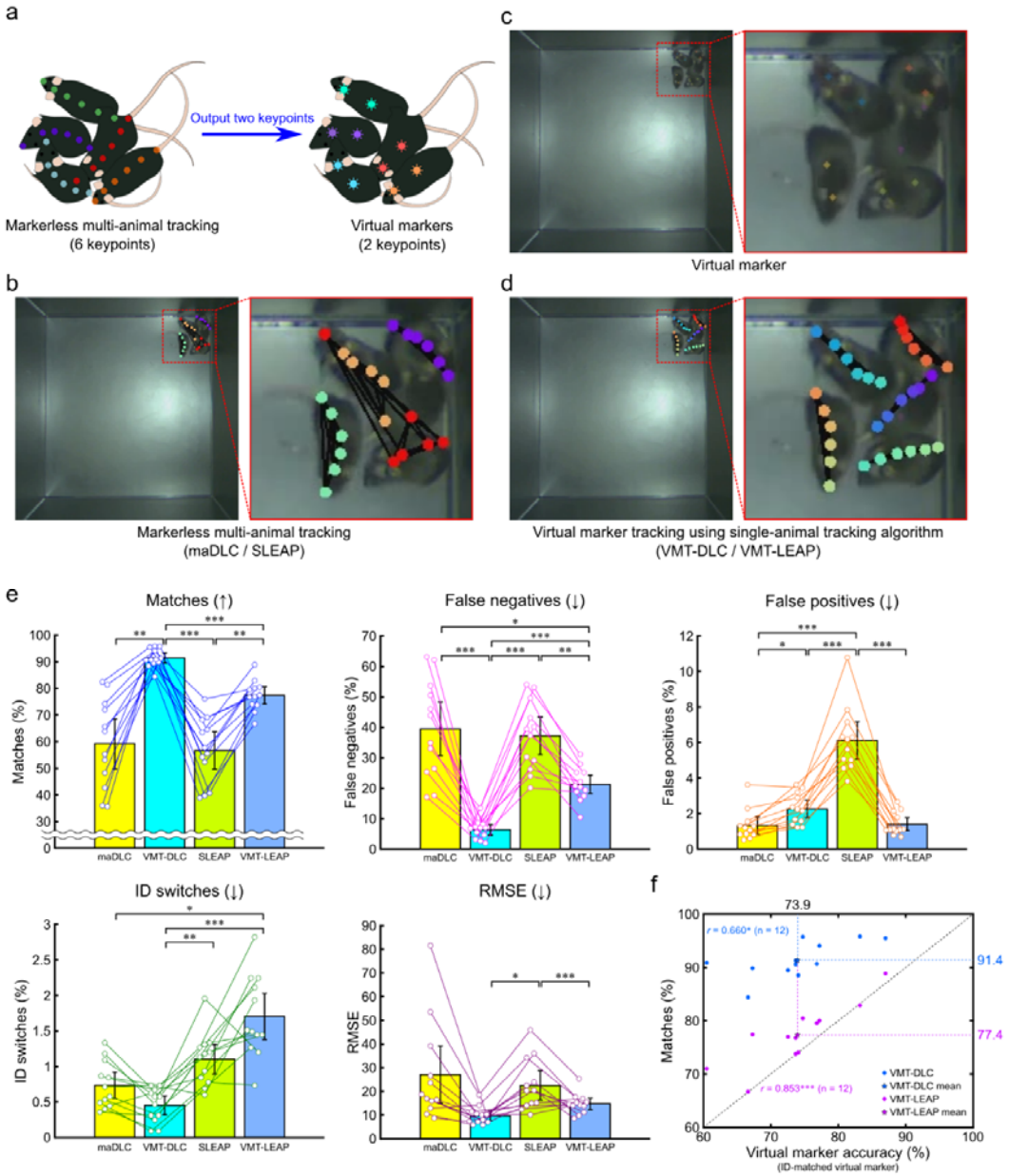

Figure 2. Comparison of markerless multi-animal tracking and vmTracking in five mice tracking.

(a) Schematic of virtual marker creation. (b–d) Snapshots of markerless multi-animal tracking in occlusion-crowded scenes (b), virtual markers (c), and vmTracking (d). The photo enclosed in the red line on the right is an enlarged view of the area enclosed in the red dotted line in the photo on the left. (e) For 12 sampled occlusion-crowded scenes (1800 frames in total), the matching rate (Matches), false negative rate (False negatives), false positive rate (False positives), ID switch rate

(ID switches), and root mean square error (RMSE) were calculated by comparing with manually created ground truth. These were compared for multi-animal DeepLabCut (maDLC), vmTracking with single-animal DeepLabCut (VMT-DLC), Social LEAP (SLEAP), and vmTracking with LEAP (VMT-LEAP). Arrows indicate whether higher or lower values are better. The results stem from five iterations of retraining for both maDLC and VMT-DLC, with the final total number of annotated frames being 1097 and 1196, respectively. The outcomes for SLEAP and VMT-LEAP are based on training and predictions by the import of annotation data from maDLC and VMT-DLC. Data represent mean ± 95% confidence interval ($n = 12$ scenes). Statistical tests were performed using repeated-measures ANOVA and Bonferroni correction, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. (f) The scatter plots illustrating the accuracy of the virtual markers based on ID matching rate, and the corresponding matching rates in vmTracking, are shown for VMT-DLC and VMT-LEAP. The filled circles represent values for individual scenes, and the stars signify their mean values. Correlation analysis was conducted, $n = 12$ scenes, *: $p < 0.05$, ***: $p < 0.001$.

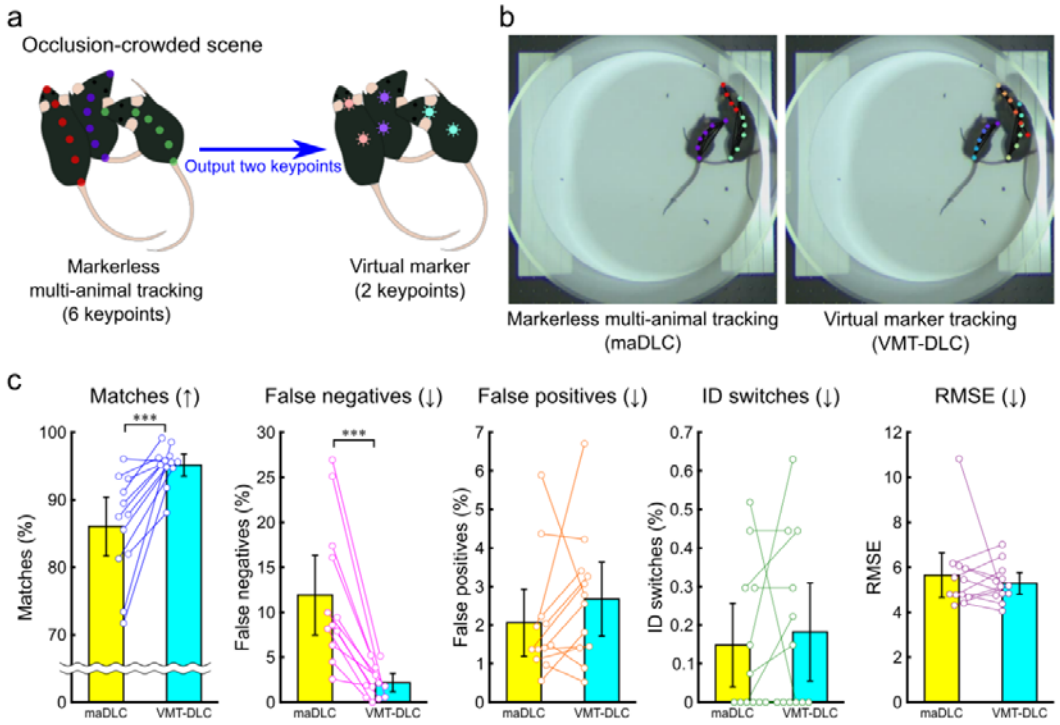

Figure 3. Comparison of markerless multi-animal tracking and vmTracking in an occlusion-crowded scene of three mice.

(a) Schematic of virtual marker creation. (b) Snapshot examples of markerless multi-animal tracking in an occlusion-crowded scene (left) and vmTracking (right). (c) For 12 sampled occlusion-crowded scenes (1950 frames in total), keypoints were compared with manually created ground truths, and the matching rates (Matches), false negative rates (False negatives), false positive rates (False positives), ID switch rates (ID switches), and root mean square errors (RMSE) were calculated. Multi-animal DeepLabCut (maDLC) and vmTracking with single-animal DeepLabCut (VMT-DLC) were compared. Arrows indicate whether higher or lower values are better. The outcomes for maDLC and VMT-DLC were obtained after 11 iterations of retraining (with a final total of 3886 annotated frames) and 5 iterations of retraining (with a final total of 1167 annotated frames), respectively. Data are shown as mean ± 95% confidence interval (n = 12 scenes). Statistical testing was performed using the Wilcoxon signed-rank test, ***: $p < 0.001$.

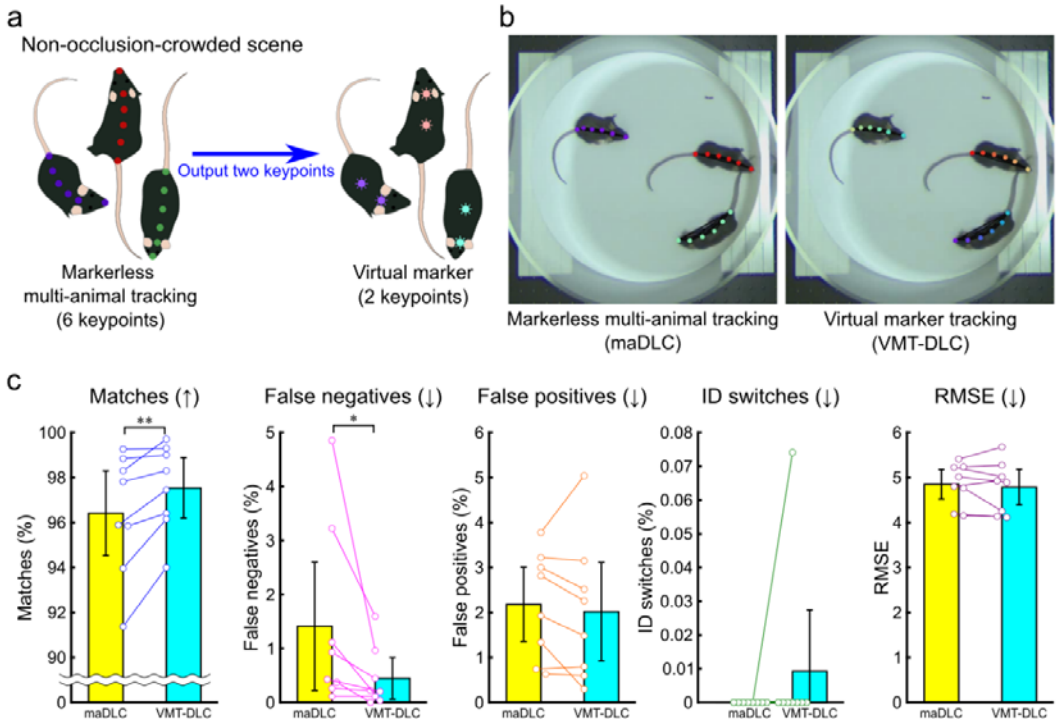

Figure 4. Comparison of markerless multi-animal tracking and vmTracking in a non-occlusion-crowded scene of three mice.

(a) Schematic of virtual marker creation. (b) Snapshot examples of markerless multi-animal tracking in a non-occlusion-crowded scene (left) and vmTracking (right). (c) For eight sampled non-occlusion-crowded scenes (1350 frames in total), keypoints were compared with manually created ground truths, and the matching rates (Matches), false negative rates (False negatives), false positive rates (False positives), ID switch rates (ID switches), and root mean square errors (RMSE) were calculated. Multi-animal DeepLabCut (maDLC) and vmTracking with single-animal DeepLabCut (VMT-DLC) were compared. Arrows indicate whether higher or lower values are better. The outcomes for maDLC and VMT-DLC were obtained after 11 iterations of retraining (with a final total of 3886 annotated frames) and 5 iterations of retraining (with a final total of 1167 annotated frames), respectively. Data are shown as mean ± 95% confidence interval (n = 8 scenes). Statistical testing was performed using the Wilcoxon signed-rank test, *: $p < 0.05$, **: $p < 0.01$.

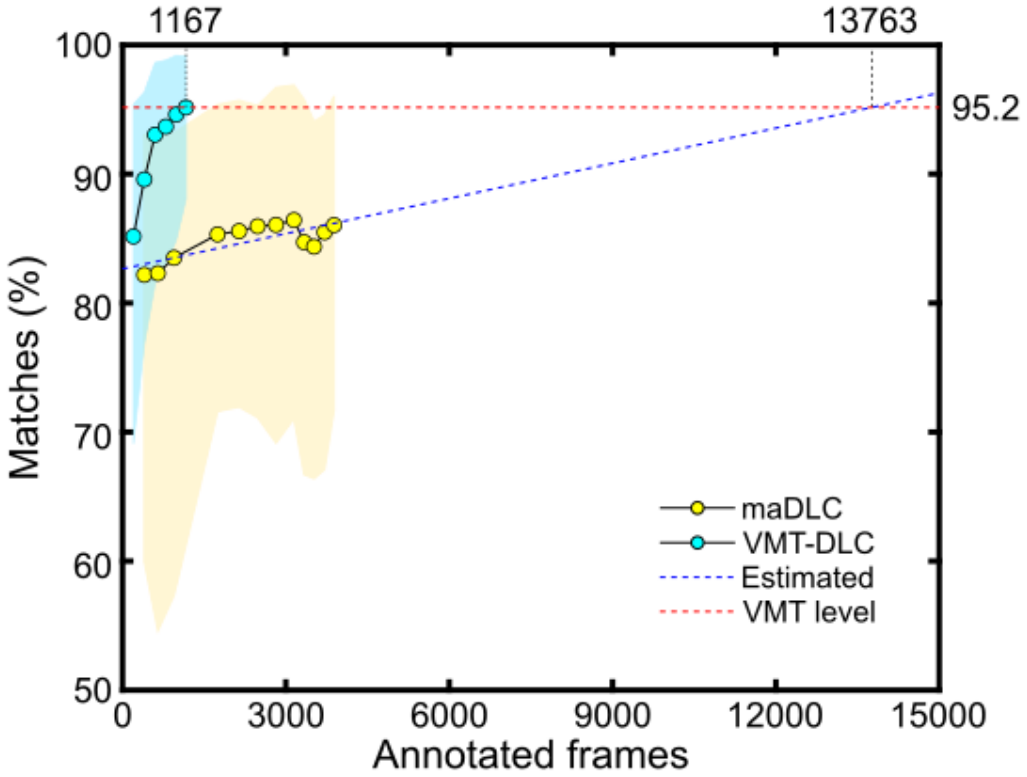

Figure 5. Estimating required annotation frames for markerless multi-animal tracking to reach an accuracy of vmTracking.

The relationship between the number of annotation frames and the average matching rate was demonstrated for both multi-animal DeepLabCut (maDLC) and vmTracking using single-animal DeepLabCut (VMT-DLC) in occlusion-crowded (OC) scenes for tracking three mice. Regression analysis using the least-squares method was conducted to estimate the number of annotation frames required for maDLC to reach the accuracy level of VMT-DLC based on the matching rate. The yellow markers represent maDLC, and the teal markers represent VMT-DLC. These plots display the number of training iterations, with the far left indicating iteration-0. The shadows in yellow and teal indicate the ranges of minimum and maximum values, respectively. The blue dotted line represents the approximated curve drawn from the performance of maDLC, and the red dotted line indicates the maximum average matching rate level (95.1%, n = 12 scenes) for the OC scene of VMT-DLC.

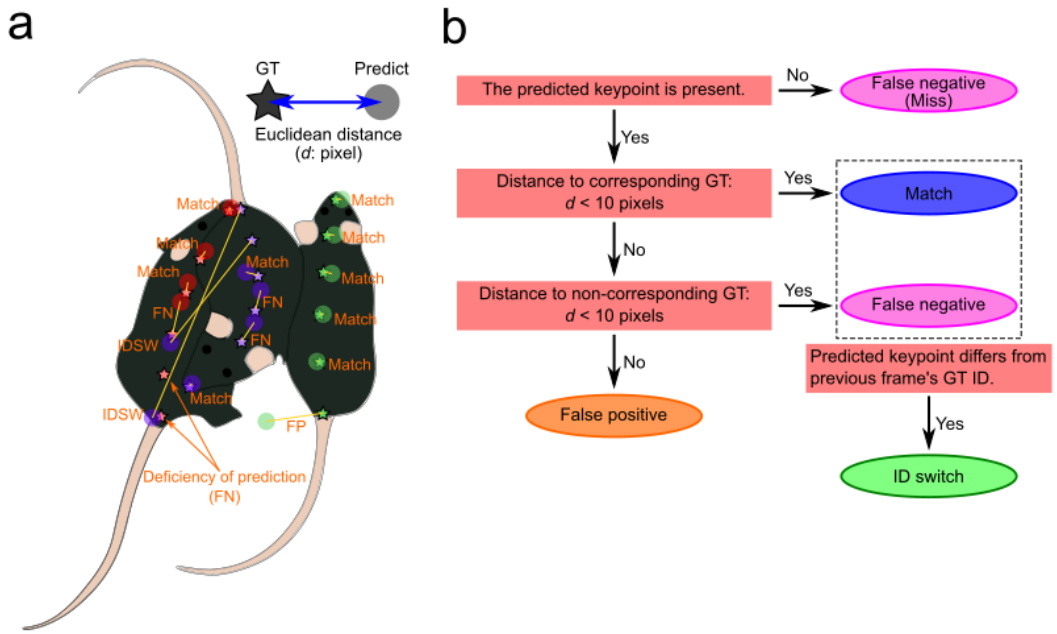

Figure S1. Classification of predicted keypoints for tracking performance evaluation.

(a) Schematic of classification of predicted keypoints. Classification into Match, false negative (FN), false positive (FP), and ID switch (IDSW) was based on the distance (*d*, in pixels) between the predicted keypoints indicated by circles and the manually created ground truth indicated by stars. The yellow lines connect the corresponding predicted keypoints to the ground truth. (b) Workflow of predicted keypoint classification. If no predicted keypoint was present, it was considered a miss and treated as an FN. If a predicted keypoint was within 10 pixels of the corresponding ground truth, it was considered a match. If a predicted keypoint was within 10 pixels of a non-corresponding ground truth, it was considered an FN because it did not predict the correct point. In cases where both the Match and FN criteria were met, the match was prioritized. If it was not within 10 pixels of any ground truth, it was considered an FP, predicting a non-existent point. In addition, if the ID determination of a predicted keypoint differed from the ID determination of the previous frame, then an ID switch occurred.

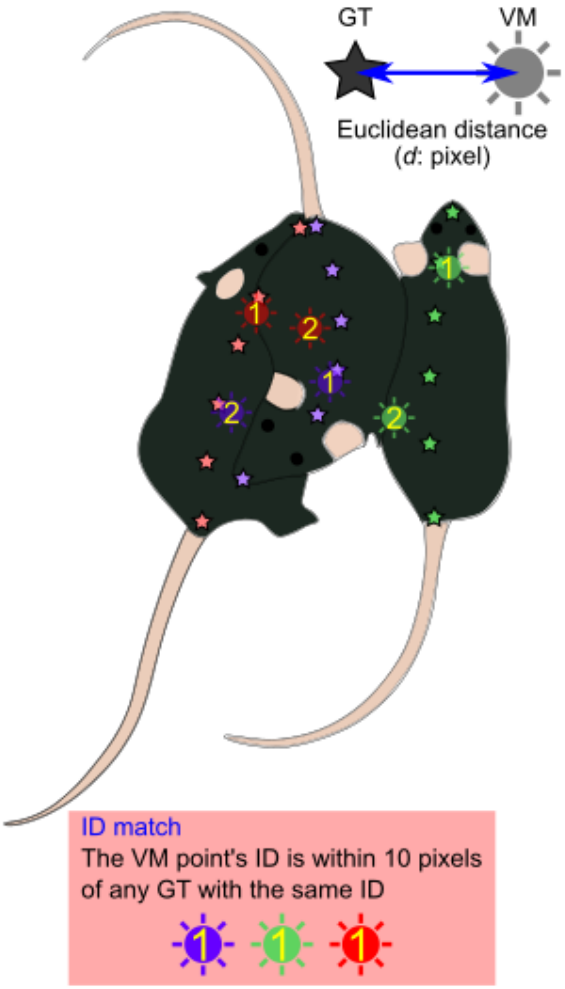

Figure S2. Accuracy of virtual markers based on ID matching.

The accuracy of the virtual markers was based on the distance (*d*: pixels) between the virtual marker points indicated by radial circles and the ground truth indicated by stars. If any of the ground truth keypoints (displayed in the same color as the virtual marker point in the figure) with the same ID as the virtual marker point were within 10 pixels of the corresponding ground truth, the virtual marker was considered to have an ID match. Number 1 in each color, indicated by a radial circle, corresponds to this. The proportion of ID match virtual marker points for each scene was calculated and used for the correlation analysis with the accuracy of vmTracking. The virtual markers utilized keypoints 2 and 4 from the results of markerless multi-animal tracking. Therefore, if these keypoints were missing, the corresponding virtual markers were also absent. The accuracy of the virtual markers calculated these missing virtual markers as instances in which there was a failure in ID matching.

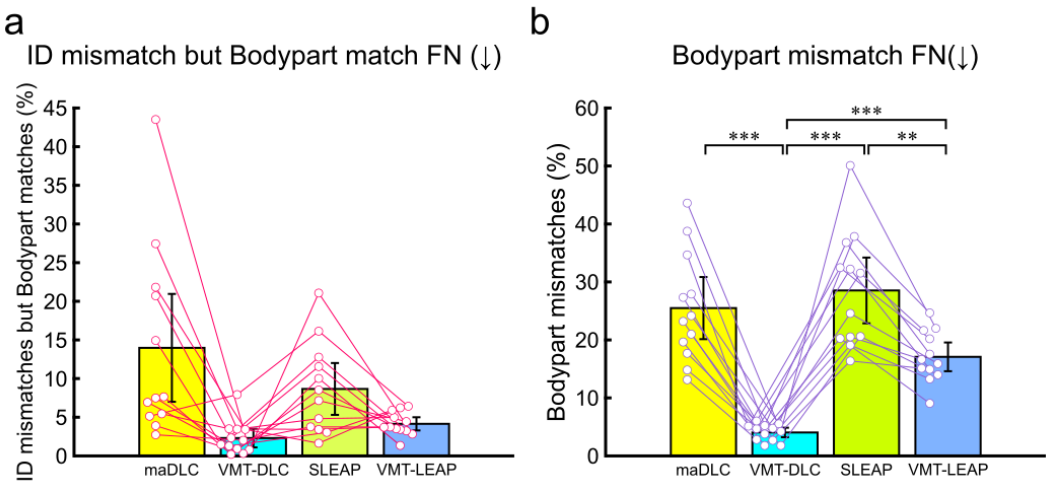

Figure S3. Evaluation through detailed classification of false negatives in five mice tracking. (a–b) For the 12 occlusion-congested scenes sampled (1800 frames in total), a comparison was made with the manually created ground truths. This comparison examined two types of false negatives: (a) cases where the predicted ID does not match, but the body parts do match (ID mismatch but Bodypart match), and (b) cases where the predicted body parts do not match, including missing parts (Bodypart mismatch). These were compared for multi-animal DeepLabCut (maDLC), vmTracking with single-animal DeepLabCut (VMT-DLC), Social LEAP (SLEAP), and vmTracking with LEAP (VMT-LEAP). Arrows indicate whether higher or lower values are better. Data represent mean ± 95% confidence interval (n = 12 scenes). Statistical tests were performed using repeated-measures ANOVA and Bonferroni correction, **: $p < 0.01$, ***: $p < 0.001$.

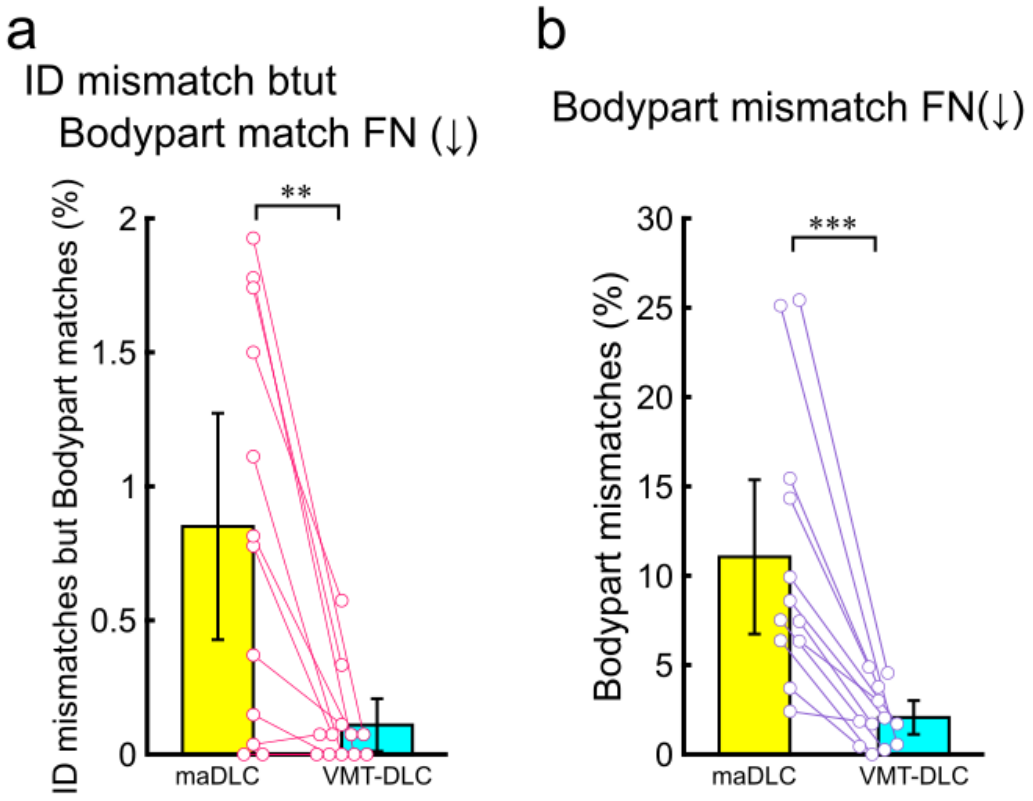

Figure S4. Evaluation through detailed classification of false negatives in an occlusion-crowded scene of three mice.

(a–b) For the 12 occlusion-congested scenes sampled (1950 frames in total), a comparison was made with the manually created ground truths. This comparison examined two types of false negatives: (a) cases where the predicted ID does not match, but the body parts do match (ID mismatch but Bodypart match), and (b) cases where the predicted body parts do not match, including missing parts (Bodypart mismatch). Multi-animal DeepLabCut (maDLC) and vmTracking with single-animal DeepLabCut (VMT-DLC) were compared. Arrows indicate whether higher or lower values are better. Data are shown as mean ± 95% confidence interval (n = 12 scenes). Statistical testing was performed using the Wilcoxon signed-rank test, **: $p < 0.01$, ***: $p < 0.001$.

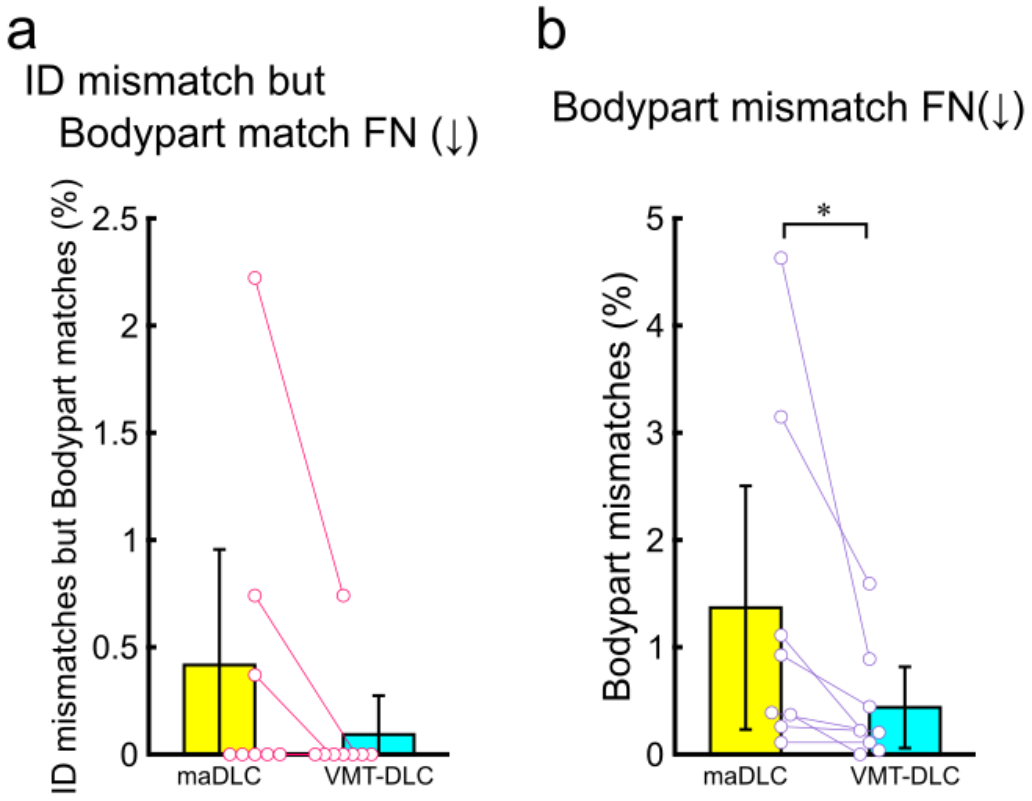

Figure S5. Evaluation through detailed classification of false negatives in a non-occlusion-crowded scene of three mice.

(a–b) For eight sampled non-occlusion-crowded scenes (1350 frames in total), a comparison was made with the manually created ground truths. This comparison examined two types of false negatives: (a) cases where the predicted ID does not match, but the body parts do match (ID mismatch but Bodypart match), and (b) cases where the predicted body parts do not match, including missing parts (Bodypart mismatch). Multi-animal DeepLabCut (maDLC) and vmTracking with single-animal DeepLabCut (VMT-DLC) were compared. Arrows indicate whether higher or lower values are better. Data are shown as mean ± 95% confidence interval (n = 8 scenes). Statistical testing was performed using the Wilcoxon signed-rank test, *: $p < 0.05$.

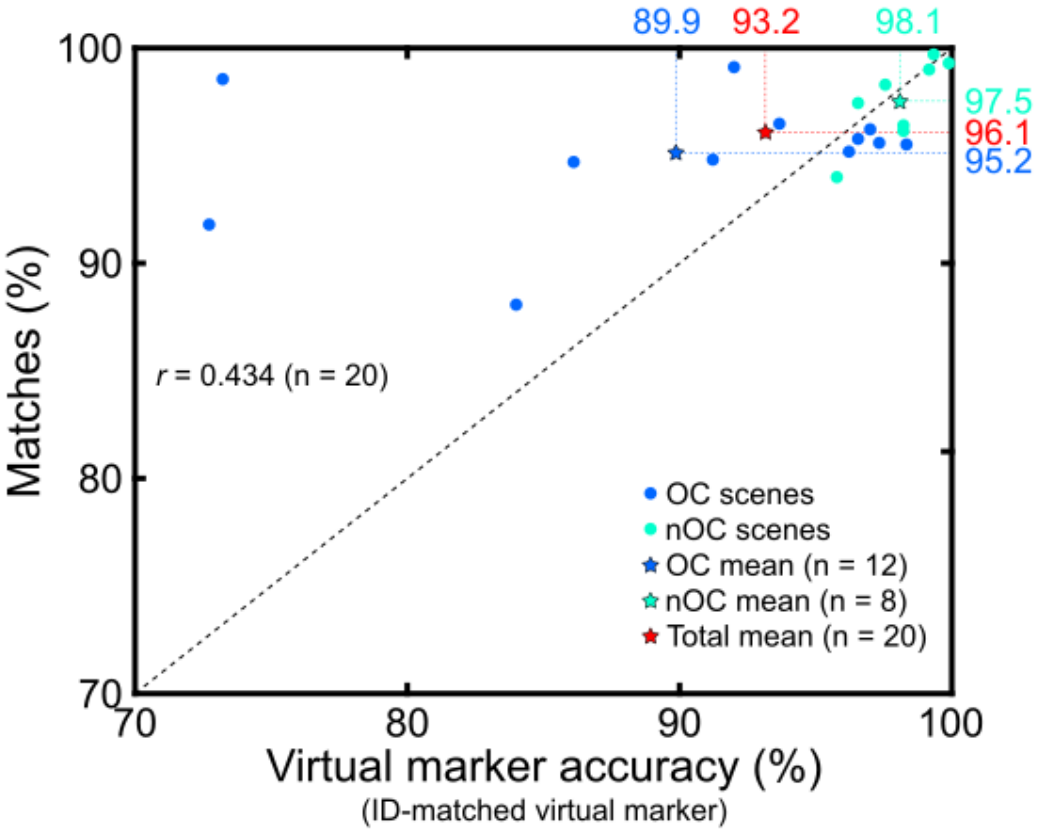

Figure S6. Correlation between the accuracy of virtual markers and the matching rate of VMT-DLC in three mice tracking.

This scatter plot shows the accuracy of the virtual markers based on the ID matching rate and the corresponding concordance rate of vmTracking using single-animal DeepLabCut (VMT-DLC). The scenes were differentiated by color into occlusion-crowded (OC) and nonocclusion-crowded (nOC) scenes. The filled circles represent the values for each sampling scene, and the stars indicate their mean values. Correlation analyses were conducted: OC, n = 12 scenes; nOC, n = 8 scenes; total, n = 20 scenes.

Supplementary information

Supplementary Video 1

The comparative video clip demonstrates the performance of maDLC (A) and VMT-DLC (B) during the tracking of five mice, with the VMT-DLC video having virtual markers excluded. In the maDLC video, at approximately 7 to 8 seconds, an identity switch occurs between the mice labeled in red and orange. Furthermore, around 1 minute and 30 seconds, during a scene of occlusion and crowding, even more complex identity switches are observed. These scenes also often involve missing predicted labels. In contrast, VMT-DLC consistently maintains accurate tracking throughout.

Supplementary Video 2

The comparative video clip demonstrates the performance of maDLC (A) and VMT-DLC (B) during the tracking of three mice in an occlusion-crowded scene. The video of VMT-DLC has virtual markers excluded. Beyond the five-second mark in the video, the mouse labeled in red is partially obscured by the mouse labeled in green. In this scene, while the prediction of the mouse labeled in red is almost missing in maDLC, VMT-DLC maintains reasonable tracking.

Supplementary Video 3

The comparative video clip demonstrates the performance of maDLC (A) and VMT-DLC (B) during the tracking of three mice in a non-occlusion-crowded scene. The video of VMT-DLC has virtual markers excluded. In scenes without occlusion or crowding, both systems maintain consistent and accurate tracking.