Upper Bound Tracker: A Multi-Animal Tracking Solution for Closed Laboratory Settings

Alexander Dolokov[1,*][a], Niek Andresen[1,3,*][b], Katharina Hohlbaum[4][c],
Christa Thöne-Reineke[2,3][d], Lars Lewejohann[2,3,4][e] and Olaf Hellwich[1,3][f]

[1]*Department of Computer Vision & Remote Sensing, Technische Universität Berlin, 10587 Berlin, Germany*

[2]*Institute of Animal Welfare, Animal Behavior, and Laboratory Animal Science, Department of Veterinary Medicine, Freie Universität Berlin, 14163 Berlin, Germany*

[3]*Science of Intelligence, Research Cluster of Excellence, Marchstr. 23, 10587 Berlin, Germany*

[4]*German Federal Institute for Risk Assessment (BfR), German Centre for the Protection of Laboratory Animals (Bf3R),*

https://www.scienceofintelligence.de

Keywords: Multiple Object Tracking, Upper Bound Tracker, Identity Switches, Mouse Home Cage Surveillance.

Abstract: When tracking multiple identical objects or animals in video, many erroneous results are implausible right away, because they ignore a fundamental truth about the scene. Often the number of visible targets is bounded. This work introduces a multiple object pose estimation solution for the case that this upper bound is known. It dismisses all detections that would exceed the maximally permitted number and is able to re-identify an individual after an extended period of occlusion including the re-appearance in a different place. An example dataset with four freely interacting laboratory mice is additionally introduced and the tracker's performance demonstrated on it. The dataset contains various conditions ranging from almost no opportunity to hide for the mice to a fairly cluttered environment. The approach is able to significantly reduce the occurrences of identity switches - the error when a known individual is suddenly identified as a different one - compared to other current solutions.

1 INTRODUCTION

Automatic video analysis often requires tracking of specific objects in the scene. That means a computer system has to be able to recognize and localize something, which it has been told to follow, in every frame of a video. In the application to observing animals there can be the additional requirement to track not only one individual and its body parts, but multiple simultaneously. To the human observer individuals can appear identical, while - through the utilization of visual appearance and the time component - the system has to be able to distinguish and identify them.

1.1 Multiple Object Tracking and Pose Estimation

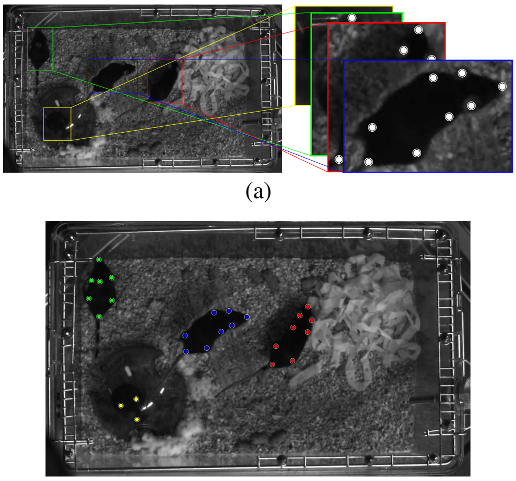
Multiple Object Tracking (MOT) is challenging and solutions are often not good enough without human correction of error. In this work, we consider the case, where a number of nearly identical individuals and their pre-defined (body) parts should be tracked across all frames (Multi-Object Pose Estimation). Our contribution is not limited to the task of pose estimation, but can also be used for situations where no keypoints play a role. Since it is most useful in laboratory animal settings, in which pose is often necessary, we present it in the Multi-Object Pose Estimation context.

1.2 Typical Frameworks

A typical Multi-Object Pose Estimation framework performs three steps (top-down, Figure 1 (a)): 1) Object Detection, 2) Body Part Detection and 3) Track-

[a] https://orcid.org/0000-0003-0207-4372
[b] https://orcid.org/0000-0002-3596-0795
[c] https://orcid.org/0000-0001-6681-9367
[d] https://orcid.org/0000-0003-0782-2755
[e] https://orcid.org/0000-0002-0202-4351
[f] https://orcid.org/0000-0002-2871-9266
[*]These authors contributed equally to this work

Dolokov, A., Andresen, N., Hohlbaum, K., Thöne-Reineke, C., Lewejohann, L. and Hellwich, O.
Upper Bound Tracker: A Multi-Animal Tracking Solution for Closed Laboratory Settings.
DOI: 10.5220/0011609500003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 5: VISAPP*, pages 945-952
ISBN: 978-989-758-634-7; ISSN: 2184-4321



(a)

(b)

Figure 1: (a) Top-down processing: An object detector finds the individuals. Another detector finds the body parts around the location of the detected objects. (b) Bottom-up processing: All body parts on the whole image are detected. As a separate step they have to be assembled and thereby assigned to individuals.

ing. Object Detection finds the individuals, Body Part Detection finds the body parts of each individual and Tracking assigns every detection to an individual. In contrast to detecting all occurring body parts on the whole image (bottom-up, Figure 1 (b)), the top-down structure allows a better resolution when detecting body parts, because the detection is run on a crop of the original image. On the other hand it necessitates the training of two separate networks: the object detector and the body part detector.

1.3 Other Tracking Solutions

There are two recent tracking solutions attacking the problem from slightly different angles. **DeepLabCut** (DLC) by Mathis et al. (Mathis et al., 2018) is an accurate tracker being used in many applications on animal videos. It is based on the first step of DeeperCut (Insafutdinov et al., 2016) - a model for human pose estimation. DLC uses a pre-trained ResNet (He et al., 2016) architecture for feature extraction followed by deconvolutions outputting a heatmap locating the specific body part. It is able to reliably find arbitrary image features based on just a few hundred training examples usually. With the recent release of version 2.2 it is also able to track multiple individuals at a time (Lauer et al., 2021). Here the authors use a different order of the steps sketched in subsection 1.2. They first perform body part detection and then assemble all the individuals (i.e. bottom-up) claiming, that the object detection as the first step often fails, when multiple individuals interact. At the end of tracking in DLC a stitching operation is performed, that optimizes the tracks globally. Each pair of consecutive tracklets gives an affinity value and the merging of tracklets is chosen such that the total affinity is minimal. Here the optimal choice is found by a min-cost flow algorithm. In contrast to the proposed method, DLC internally creates a model of how the individual bodyparts compose the whole, such that the detected part can be attributed to the right individual, even if other individuals are close by. The false detection of only one body part can trigger the creation of a new individual track - an event the proposed approach tries to prevent. **SLEAP** is another open-source tracking framework (Pereira et al., 2022). It includes both bottom-up and top-down approaches also for multiple individuals and their body parts. It relies on an interactive learning process with a human in the loop. The user labels some data, lets the method predict and then fixes erroneous detections, which are then used for further training and so on. For step 3) Tracking two options are offered: Optical Flow or Kalman Filter. Both try to generate a prediction of where a track will continue in a new frame. Those predictions are then matched to the detections minimizing the matching cost. Both DLC and SLEAP allow a manual repair of switched identities during tracking. False detections have to be removed manually, since no fixed upper bound is employed.

1.4 Multitracker Features

The **Multitracker** framework introduced in this work utilizes currently successful deep learning methods for all steps and introduces a novel approach to step 3) Tracking, that leverages the knowledge of the maximum number of individuals present, which is available in many laboratory animal applications.

For Step 1) Object Detection. the here implemented method is YOLOX (Ge et al., 2021). The YOLO approach handles the detection and classification of objects in an image in one deep network, while outperforming alternative methods (Redmon et al., 2016) such as Faster R-CNN (Ren et al., 2016). We chose YOLOX, because it combines high quality predictions with high efficiency. It allows differently scaled models, which enables users to tune the trade-off between speed and accuracy themselves. SSD (Liu et al., 2016) is another successful method, but it did not perform as well as YOLOX on the mouse data while being comparable in speed. SSD is thus not included in the Multitracker framework.

For Step 2) Keypoint Detection. the here implemented especially successful and frequently used options are Efficient U-Net (Ronneberger et al., 2015), Stacked Hourglass Network (Newell et al., 2016) and Pyramid Scene Parsing Network (PSP) (Zhao et al., 2017). These methods are available in the provided framework, but are not elaborated or evaluated in this work.

For Step 3) Tracking. four methods are offered. The two widely adopted MOT algorithms SORT (Bewley et al., 2016) and the V-IoU Tracker (Bochinski et al., 2018), the current state-of-the-art OC-SORT (Cao et al., 2022) as well as the novel Upper Bound Tracker introduced in this work. All four perform track assignment, estimation, and management based on bounding boxes created by an object detector. A motion model or the current frame is used to estimate the current location using the past track data. These estimated tracks are then matched with the new detections. Afterwards the creation and deletion of tracks is managed based on simple rules. The different approaches in each of these steps distinguish the methods. **SORT** (Bewley et al., 2016) is a tracking method, that utilizes a Kalman Filter (Kalman, 1960) to estimate the next position in a track. Afterwards it maximizes the intersection over union (IoU) between tracks and detections with the help of the Hungarian algorithm. SORT creates new tracks after an unmatched detection and deletes them if they could not be matched with a detection too many time steps in a row. This sophisticated method is able to cope with inconsistent detections through the Kalman filter. Observaion-Centric SORT (**OC-SORT**) (Cao et al., 2022) is based on SORT, but introduces improvements to the Kalman Filter step. There the predictions for the next step are not assumed linear, which leads to large improvements over SORT in situations of occlusions and non-linear movement. The Visual-Intersection-over-Union (**V-IoU**) tracker (Bochinski et al., 2018) relies on more consistent detections. A new detection is matched to a track by computing the IoU between it and the previous detections. If the intersection is high, the new detection likely belongs to that track. Unmatched tracks are continued with a visual tracker to fill detections gaps at least for some number of time steps. The same is done backwards in time with unmatched detections. The fourth and final tracking method is designed for a slightly less general setting, that is introduced in the next section.

The code is publicly available on GitHub[1].

2 UPPER BOUND TRACKING

Most MOT benchmarks (Dendorfer et al., 2020) track objects in open world settings, e.g. public surveillance cameras in public spaces. Video sequences and their corresponding tracks are relatively short. No prior information about the total number of individual objects is known. In some behavioural observation experiments however, cameras film animals within a cage. In this closed world setting, a small number of subjects is filmed for a long time. For each video, the total number of participating animals is known. We call this setting ”Upper Bound Tracking” as it contains a strict upper bound for the number of visible subjects at any time. Utilizing this knowledge can improve tracking significantly and is at the center of the proposed Upper Bound Tracker (UBT). By careful design, tracking rules can be derived that guarantee to never violate the upper bound while at the same time increase global track consistency.

3 METHOD

The Upper Bound Tracker (UBT) is based on OC-SORT (Cao et al., 2022) and contains adjustments to the creation of new tracks and to the reconnection of lost tracks. It is designed to reduce identity switches compared to other trackers by preventing spurious detections to create new tracks. Like SORT and OC-SORT, the UBT is similar to the V-IoU Tracker (Bochinski et al., 2018) by assigning a new detection to a track if they have a large IoU. But it never creates new tracks if the upper bound for the number of individuals is already reached. Additionally a novel reidentification step is introduced, that connects a previously lost track to a new appearing one. In conjunction with the strict upper bound this reidentification takes effect, when an animal was occluded for an extended period of time - leading to less than the maximum amount of individuals visible - and reappears at a later point. This way the correct identity is assigned again given that in the meantime the other individuals were not also lost from sight.

The frame update step is presented in Algorithm 1. It describes the steps, that are performed after the detections have been made on the new frame and the Kalman Filter has predicted the next bounding boxes. In the frame update step an unmatched track is set to inactive after it has not been matched with a detection for a set number of time steps ($*^1$ in Algorithm 1). An unmatched detection is matched with the closest inactive track, when it is stable ($*^2$). We call a detection stable, when in each of the last three time steps there

[1] https://github.com/dolokov/upper_bound_tracking

was a detection close by it (IoU $> \frac{1}{2}$) - i.e. it is stable, when it did not appear far away from all other recent tracks.

The described approach results in all additional detections being discarded when the upper bound is already reached. This is only correct if the existing tracks are all following the actual individuals and are not due to some spurious detections. The chance of such a fault happening are reduced by the need for detections to be stable before being attached to a track, as well as the required small distance to the last known track position. Only close-by and very continuous false detections could cause an issue, that - given the current framework and data - is only prevented by using a good object detector for step 1). When false detection occur only briefly for a few frames, they are unlikely to cause any problem for the proposed method, while other methods will create new tracks for them.

4 DATASET AND EVALUATION

We created videos to test the tracker's performance in a setting, where the Upper Bound Tracker approach might be useful in the future: videos of a fixed number of animals moving in a closed cage. In the videos four mice are freely moving through a 425 x 276 mm (type III) polycarbonate cage, that is filmed from above such that the whole cage is in the frame. The filter top as well as grid of the cage were removed and replaced by a custom-made transparent lid of the same size, which prevented the mice from climbing onto and walking along the edge of the cage walls. During video recording, food pellets normally supplied as diet (LASvendi, LAS QCDiet, Rod 16, autoclavable) were placed on the floor. Water was provided in a bottle attached to the external wall of the cage; the drinking nipple was put through a hole in the cage wall so that the mice had free access to water during the video recording. The video dataset is publicly available[2].

```
Input: $u$, $T$, $D$, $n_{ia}$, $d_{cl}$, $d_{reid}$, $n_{\text{misses}}$
Result: new Tracks $T'$
Match tracks $T$ to detections $D$ with Linear
  Programming with IoU criterion;
/* Update tracks for good matches
   */
foreach matched pair of track and detection
 (t,d) do
    update track attributes $t' \leftarrow \frac{1}{2}(t+d)$;
    $n^t_{\text{misses}} \leftarrow 0$;
    set $t$ to active;
end
/* Set lost tracks to inactive    */
foreach unmatched track $t$ do
    $n^t_{\text{misses}} \leftarrow n^t_{\text{misses}} + 1$;
    if $n^t_{\text{misses}} \geq n_{ia}$ then Set $t$ inactive;  // $*^1$
end
foreach unmatched stable detection $d$ do
    /* If there are too few tracks
       add a new one              */
    if $|T| < u$ then
        $d_{min} \leftarrow \min_{t \in T}$ dist($d$,$t$);
        if $d_{min} > d_{cl}$ then
            add new track at position $d$ to $T$
        end
    end
    /* Otherwise add detection to
       closest inactive track      */
    else
        $t_{\text{closest}} \leftarrow \text{argmin}_{t \in T_{\text{inactive}}}$ dist($d$,$t$);
        if $dist(t_{\text{closest}}, d) < d_{reid}$ then
            interpolate between the last
              matched location of $t_{\text{closest}}$ and
              $d$;                              // $*^2$
            set $t_{\text{closest}}$ to active;
        end
    end
end
```

Algorithm 1: The frame update step. Inputs are: u - the upper bound, T - the current Kalman Filter predicted locations and sizes of the tracks, D - the new detections, n_{ia} - the number of time steps after a lost track is set to inactive, d_{cl} - the minimum clearance distance, d_{reid} - the maximum reidentification distance. Detections d and tracks t consist of location (x,y), width and height.

The mice were video-recorded under ten different environmental enrichment conditions; i.e., for each video segment different enrichment items were provided to the mice - from here on called occlusion conditions or just conditions (see Table 1). The more objects were present, the more occlusions could occur. In all occlusion conditions, the cage floor was covered with wooden bedding material (JRS Lignocel FS14, spruce/ fir, 2,5-4 mm) and 5 g shredded cotton cocoons (UNIGLOVES Dental Watterollen Gr.3). In the most crowded occlusion condition, there are a transparent tunnel, a house with a running plate, some paper strips, and paper towel, which offered the mice lots of options to hide from the camera and should be challenging for any tracker. Sample frames from those two most extreme conditions can be found in Figure 2. For more information on the camera setup see section 7. This kind of data can be found in experiments observing the social life of mice. The individual has to be recognized in order to judge e.g.

[2] https://www.scienceofintelligence.de/research/data/four-mice-from-above-dataset/

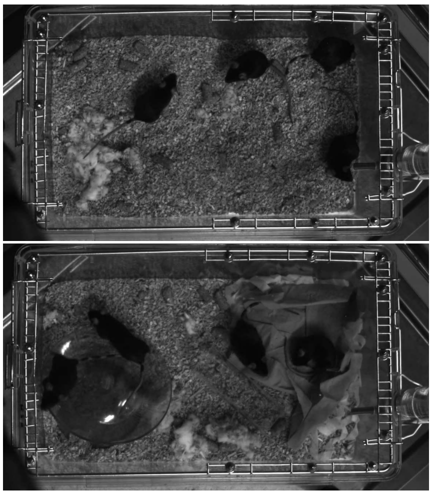


Figure 2: Example frames of the least and the most occluded conditions.

each mouse's activity level or the number of interactions with other mice. Tracking can help to do this automatically, but a high number of identity switches will dampen its usefulness. They have to be corrected manually forcing a researcher to watch the whole sequence again. Thus the number of switches has to be minimal in such an application.

For training the method the YOLOX-M (Ge et al., 2021) object detection model was trained on frames taken from four different occlusion conditions including the least and the most occluded conditions. 300 frames were taken from the beginning of each of the four video segments with a distance of 50 frames (or 1.67 seconds) between each other. The resulting 1200 frames were labeled with bounding boxes around the mice. 10% or 120 frames were taken as validation set. Training was performed until convergence (about 300 epochs). Afterwards the three tracking methods were run and their performance evaluated. For evaluation a number of video snippets were annotated manually. Every 50th frame was shown to the annotator, who then drew bounding boxes around each visible mouse and assigned the boxes to an individual. Individuals are recognizable in the videos through the markings on their tail. For the gaps of 49 frames (or 1.6 seconds) between bounding-box annotated frames the bounding boxes were interpolated. This was done for the first minute of six videos with different occlusion conditions. Note, that localization performance was not evaluated here.

These obtained ground truth tracks were used for evaluation with the HOTA metric (Higher Order Tracking Accuracy) (Luiten et al., 2021). This recently published metric balances the measurement of performance of a tracker in correct detection and correct association, while eliminating a number of shortcomings, that common metrics like MOTA (Bernardin and Stiefelhagen, 2008) and IDF1 (Ristani et al., 2016) have. For these metrics a higher value is better.

Since a good tracker in applications to laboratory animals science and elsewhere has to follow each individual reliably the number of identity switches were separately counted. Here 'identity switch' refers to the event, that an animal is assigned to a track, that was previously associated with a different animal.

A comparison is also made to the complete multiple-object pose estimation solution DeepLabCut (DLC) in version 2.2. DLC does not output bounding boxes, but the metrics HOTA and MOTA are (partially) computed with a similarity score between bounding boxes. To be able to consider these metrics as well, we determined the bounding boxes of the keypoints, that DLC outputs and increased their width and height by 10%. On those metrics the comparison is not fair, because the bounding boxes stem from an approximate heuristic, so the values are not important to consider. The other metrics are more meaningful here. We used the same training data as for the other methods and trained a multi-animal DLC model using default parameters.

The final experiment presented here delivers evidence that the introduction of the upper bound leads to better results. To this end the method is applied to the same data, but with an upper bound that is too high.

5 RESULTS

In the following comparisons between the UBT and the aforementioned approaches for step 3) Tracking (V-IoU, SORT and OC-SORT trackers) are presented.

On all videos regardless of occluding objects in the scene the UBT outperforms OC-SORT, SORT and V-IoU on the metric counting the number of ID switches (IDSW). Here the difference in performance to the second best, OC-SORT, is rather small, while the difference to the other methods is substantial, cutting the number of switches in half at least.

On the other metrics it shows good performance as well. Table 2 (upper panel) shows results for the easiest condition, in which no obstacles obscure the mice. UBT performs slightly better than OC-SORT in all metrics. The HOTA, IDF1 and IDSW performance sees a big gab between the two on one side and SORT and V-IoU on the other. The MOTA score is similar for all four.

Table 1: Objects in the cage in each of the six occlusion conditions. An 'X' marks the presence of the object. Each condition has wooden bedding material and shredded cotton and can also have: tunnel (transparent, 11,5 cm x 3,5 cm, custom-made), one or two grams of white paper strips (LILLICO, Biotechnology Paper Wool), thin paper towels (cellulose, unbleached, layers, 20x20cm, Lohmann & Rauscher), Mouse igloo with or without running plate (ZOONLAB GmbH, Castrop-Rauxel, Germany; round house: 105 mm in diameter, 55 mm in height; round plate: 150 mm in diameter).

Condition	Tunnel	Igloo	Paper strips	Running Plate	Paper towels
1					
2	X				
3	X	X			
4	X	X	1g		
5	X	X	2g	X	
6	X	X	2g	X	X

Table 2: MOT performance of the five compared methods on the easiest and on the most difficult occlusion condition. HOTA: area under the curve for HOTA$_\alpha$ for α ranging from 0.05 to 0.95 in steps of 0.05; IDSW: number of identity switches; bolt: best value for each column; *: DLC could not be fairly evaluated with HOTA and MOTA (see section 4).

Easiest Occlusion Condition

	HOTA	MOTA	IDF1	IDSW
SORT	0.39	0.86	0.42	25
OC-SORT	0.56	0.85	0.74	5
V-IoU	0.33	0.84	0.33	42
UpperBound	**0.58**	**0.88**	**0.77**	4
DLC	0.54*	0.42*	0.71	**0**

Most Difficult Occlusion Condition

	HOTA	MOTA	IDF1	IDSW
SORT	0.30	**0.73**	0.31	57
OC-SORT	**0.34**	0.69	0.41	25
V-IoU	0.30	0.71	0.30	71
UpperBound	0.33	0.54	**0.48**	**22**
DLC	0.19*	-0.06*	0.27	66

The most difficult occlusion condition (Table 2 lower panel) sees OC-SORT slightly ahead of UBT in the HOTA and SORT ahead in the MOTA score. Here the OBT performs best only in IDF1 and IDSW.

Performance on the other conditions can be found in section 7 in the appendix.

DLC performs as well as OC-SORT and UBT on the least occluded condition[3]. On the most difficult condition its performance falls off, however. Here it is similar to SORT and V-IoU again.

When setting the upper bound too high performance on all metrics drops (Table 3).

Table 3: MOT performance of the UBT when setting the upper bound incorrectly. The correct upper bound for the data is 4. Evaluation was done on the easiest occlusion condition. HOTA: area under the curve for HOTA$_\alpha$ for α ranging from 0.05 to 0.95 in steps of 0.05; IDSW: number of identity switches.

Upper Bound	HOTA	MOTA	IDF1	IDSW
4	0.58	0.88	0.77	4
5	0.51	0.64	0.68	5
10	0.38	-0.41	0.46	9

6 DISCUSSION

The HOTA and IDF1 metrics have a range between 0 (nothing was done right) to 1 (perfect performance). MOTA is unbounded in the negative direction and also has an upper bound of 1. The number of ID switches can of course be any non-negative integer. This metric is dominated by the UBT with OC-SORT closely following. The good performance of it in the domain of getting the identity of the individuals right is still visible in the IDF1 metric, which has a bias towards that component to MOT performance (Luiten et al., 2021). Here the UBT again outperforms other methods by a good margin. This indicates, that the improvements, that OC-SORT and UBT brought, were mainly to the consistency of individual identification, and less to the localization accuracy. The poorer performance of UBT in the MOTA metric on the most challenging condition points towards a weakness in correctly drawing bounding boxes around individuals, that are only partially visible. The other occlusion conditions paint a similar picture. The effect of the introduction of the upper bound on the number of spurious detections becomes obvious in the ablation experiment. When setting the upper bound to ten instead of the correct four, the MOTA score even becomes negative, which happens, when often more false positives occur than there are ground truth tracks.

7 CONCLUSION

The UpperBound Tracker shows great improvements on existing baseline methods for MOT. It is also able

[3] Only considering IDF1 and IDSW - see section 4

to out-perform the recent state-of-the-art tracker OC-SORT by a small margin. The most balanced metric HOTA, which gives appropriate weight to both subtasks: finding the individuals and consistently identifying them, still shows room for improvement under challenging conditions. The other metrics give evidence, in which sub-task the contribution of the UBT idea lies. The number of identity switches is much lower. This is indicating, that the correct and consistent identification of tracked individuals benefits from the re-connection to the closest inactive track, that is introduces in the UBT in this work. Further research should address the case when more than one individual is gone from view. The reidentification could take into account past trajectories and appearances of missing tracks to connect them once they reappear. In the MOT sub-task of following and re-identifying individuals in videos, that fulfill the requirement of a known maximum number of individuals, UBT is a good choice.

ACKNOWLEDGEMENTS

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2002/1 "Science of Intelligence" – project number 390523135.

We thank Clara Bekemeier and Sophia Meier for manual data annotation and Benjamin Lang for building the transparent cage lid.

REFERENCES

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

Bewley, A., Ge, Z., Ott, L., Ramos, F., and Upcroft, B. (2016). Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE.

Bochinski, E., Senst, T., and Sikora, T. (2018). Extending iou based multi-object tracking by visual information. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE.

Cao, J., Weng, X., Khirodkar, R., Pang, J., and Kitani, K. (2022). Observation-centric sort: Rethinking sort for robust multi-object tracking. *arXiv preprint arXiv:2203.14360*.

Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K., and Leal-Taixé, L. (2020). Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv:2003.09003[cs]*. arXiv: 2003.09003.

FELASA Working Group on Revision of Guidelines for Health Monitoring of Rodents and Rabbits, Mähler, M., Berard, M., Feinstein, R., Gallagher, A., Illgen-Wilcke, B., Pritchett-Corning, K., and Raspa, M. (2014). Felasa recommendations for the health monitoring of mouse, rat, hamster, guinea pig and rabbit colonies in breeding and experimental units. *Laboratory animals*, 48(3):178–192.

Ge, Z., Liu, S., Wang, F., Li, Z., and Sun, J. (2021). Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., and Schiele, B. (2016). Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*, pages 34–50. Springer.

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45.

Lauer, J., Zhou, M., Ye, S., Menegas, W., Nath, T., Rahman, M. M., Di Santo, V., Soberanes, D., Feng, G., Murthy, V. N., et al. (2021). Multi-animal pose estimation and tracking with deeplabcut. *bioRxiv*.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer.

Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., and Leibe, B. (2021). Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129(2):548–578.

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., and Bethge, M. (2018). Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 21(9):1281–1289.

Newell, A., Yang, K., and Deng, J. (2016). Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer.

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., Shaevitz, J. W., and Murthy, M. (2022). Sleap: A deep learning system for multi-animal pose tracking. *Nature Methods*.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.

Ren, S., He, K., Girshick, R., and Sun, J. (2016). Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149.

Ristani, E., Solera, F., Zou, R., Cucchiara, R., and Tomasi, C. (2016). Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890.

APPENDIX

Ethics Statement

Maintenance of mice and all animal experimentation was approved by the Berlin State Authority and the Ethics committee (“Landesamt für Gesundheit und Soziales”, permit number: G0249/19). The study was performed according to the German Animal Welfare Act, and the Directive 2010/63/EU for the protection of animals used for scientific purposes.

Animals

Four female C57BL/6J mice obtained from Charles River Laboratories (Sulzfeld, Germany) were used at an age of approximately 10 months. The animals were group-housed in two polycarbonate type 3 cages (425 x 276 mm each) with filter tops, which were connected with each other via a tube. The cages contained wooden bedding material (JRS Lignocel FS14, spruce/ fir, 2,5-4 mm), a triangular plastic house (140 mm long side, 100 mm short sides, 50 mm in height; Tecniplast, Italy), a transparent tunnel (11 mm x 40 mm , custom-made), and five pieces of paper towel (2 x Paper Towels 23x24,8cm folded, Essity ZZ Towel; 3 x cellulose, unbleached, layers, 20x20cm, Lohmann & Rauscher). The animals were maintained under standard conditions (room temperature: 22 ± 2 °C; relative humidity: 55 ± 10 %) on a light:dark cycle of 12:12 h of artificial light (lights on from 7AM to 7PM in the winter and 8AM to 8PM in the summer) with a 30 min twilight transition phase. They had free access to water and were fed pelleted mouse diet ad libitum (LASvendi, LAS QCDiet, Rod 16, autoclavable). Cages were cleaned once a week and the mice were handled using a tunnel. The experimenter was female. The mice were free of all viral, bacterial, and parasitic pathogens listed in the FELASA recommendations (FELASA Working Group on Revision of Guidelines for Health Monitoring of Rodents and Rabbits).

Camera Setup

The video recording was done with a Basler acA1920-40um camera (Lens LM25HC F1.4 f25mm, Kowa, Nagoya, Japan) mounted on a tripod pointing down at the type III cage (425 mm × 276 mm × 150 mm) with transparent lid. The camera has a resolution of 1920 x 1200 pixels and was set to record 30 monochrome frames per second with a pixel bit depth of 8 bit.

Performance on Other Occlusion Conditions

Table 4: MOT performance of the five compared methods on different occlusion conditions. HOTA: area under the curve for HOTA$_\alpha$ for α ranging from 0.05 to 0.95 in steps of 0.05; IDSW: number of identity switches; bolt: best value for each column; *: DLC could not be fairly evaluated with HOTA and MOTA (see section 4) and was not evaluated for all conidtions.

Occlusion Condition Difficulty 2/6

	HOTA	MOTA	IDF1	IDSW
SORT	0.43	**0.77**	0.52	15
OC-SORT	**0.52**	0.72	0.67	9
V-IoU	0.40	0.74	0.48	27
UpperBound	**0.52**	0.73	**0.72**	**5**
DLC	0.46*	0.41*	0.66	19

Occlusion Condition Difficulty 3/6

	HOTA	MOTA	IDF1	IDSW
SORT	0.26	**0.62**	0.24	83
OC-SORT	0.25	0.46	0.27	55
V-IoU	0.22	0.57	0.21	118
UpperBound	**0.38**	0.53	**0.54**	**30**

Occlusion Condition Difficulty 4/6

	HOTA	MOTA	IDF1	IDSW
SORT	0.45	0.84	0.49	23
OC-SORT	0.59	0.79	0.71	3
V-IoU	0.42	0.82	0.45	33
UpperBound	**0.70**	**0.85**	**0.92**	**0**

Occlusion Condition Difficulty 5/6

	HOTA	MOTA	IDF1	IDSW
SORT	0.40	**0.61**	0.48	35
OC-SORT	0.43	0.52	0.56	36
V-IoU	0.38	0.57	0.45	53
UpperBound	**0.54**	0.57	**0.78**	**17**
DLC	0.37*	0.24*	0.51	20