Leveraging Deep Learning for Approaching Automated Pre-Clinical Rodent Models

Carl Sandelius[1,2], Athanasios Pappas[2], Arezoo Sarkheyli-Hägele[1][a], Andreas Heuer[2][b] and Magnus Johnsson[3][c]

[1] *Internet of Things and People Research Center, Department of Computer Science and Media Technology, Malmö University, Malmö, Sweden*

[2] *Behavioural Neuroscience Laboratory, Department of Experimental Medical Sciences, Lund University, Lund, Sweden*

[3] *Research Environment of Computer Science (RECS), Kristianstad University, Sweden*

Keywords: Deep Learning, Machine Learning, Computer Vision, Behavioral Neuroscience, Pre-Clinical Rodent Models.

Abstract: We evaluate deep learning architectures for rat pose estimation using a six-camera system, focusing on ResNet and EfficientNet across various depths and augmentation techniques. Among the configurations tested, ResNet 152 with default augmentation provided the best performance when employing a multi-perspective network approach in the controlled experimental setup. It reached a Root Mean Squared Error (RMSE) of 8.74, 8.78, and 9.72 pixels for the different angles. The utilization of data augmentation revealed that less altering yields better performance. We propose potential areas for future research, including further refinement of model configurations, more in-depth investigation of inference speeds, and the possibility of transferring network weights to study other species, such as mice. The findings underscore the potential for deep learning solutions to advance preclinical research in behavioral neuroscience. We suggest building on this research to introduce behavioral recognition based on a 3D movement reconstruction, particularly emphasizing the motoric aspects of neurodegenerative diseases. This will allow for the correlation of observable behaviors with neuronal activity, contributing to a better understanding of the brain and aiding in developing new therapeutic strategies.

1 INTRODUCTION

Integrating data science and machine learning techniques into neuroscientific research represents an evolving field that brings huge potential for improved efficiency gains in conducting preclinical medical trials, minimizing the influence of human bias, and introducing more quantifiable, reproducible, and scalable methodologies. This study has positioned itself at the intersection of these disciplines, focusing on taking another step toward automating preclinical rodent models in degenerative diseases.

Preclinical rodent models are imperative before commencing with human clinical trials to evaluate novel therapeutic options to halt or reverse disease. Traditionally, this process is manual, time-consuming, monotonous, and error-prone, with low inter-rater reliability and coarse rating scales.

Neurodegenerative disorders, including Parkinson's disease and Huntington's disease, along with neurological disorders such as Epilepsy, manifest through motor impairments, presenting significant challenges in their study and treatment. Traditional methods, such as manual scoring or marker-based systems, suffer numerous drawbacks that stifle research progress. Manual scoring is labor-intensive and introduces subjectivity, while marker-based systems can disrupt natural rodent behavior.

The advancements in computer vision and deep learning have transformed the analysis of rodent behavior, offering markerless capabilities that avoid the pitfalls of previous approaches. The introduction of tools such as DeepLabCut (Mathis et al., 2018), Deepfly (Günel et al., 2019), and JAABA (Kabra et al., 2013) has facilitated the application of deep neural networks in behavioral neuroscience, allowing for non-invasive, accurate tracking in video feeds. Yet the field continues to seek enhancements in 3D move-

[a] https://orcid.org/0000-0001-6925-0444

[b] https://orcid.org/0000-0003-0300-7606

[c] https://orcid.org/0000-0002-4409-1413

Sandelius, C., Pappas, A., Sarkheyli-Hägele, A., Heuer, A. and Johnsson, M.
Leveraging Deep Learning for Approaching Automated Pre-Clinical Rodent Models.
DOI: 10.5220/0013065600003837

In *Proceedings of the 16th International Joint Conference on Computational Intelligence (IJCCI 2024)*, pages 613-620
ISBN: 978-989-758-721-4; ISSN: 2184-3236

ment analysis, which can potentially integrate behavioral data with neuronal activity measures to establish causative links between brain function and behavior.

While significant strides have been made in 3D pose estimation for animal behavior analysis (Karashchuk et al., 2021) (Günel et al., 2019) (Mathis et al., 2018), the focus has predominantly been on flies and mice, with limited exploration in rats, which play a central role in studying more nuanced behavioral components in degenerative diseases. Unlike prior studies, such as the one by Nilsson et al. (Nilsson et al., 2020), which examined social behaviors in rats, this study has positioned itself to address the motoric components of seizures tied to specific degenerative conditions, tracking a more complex set of poses than before. Distinct in their behaviors compared to mice, rats are indispensable for certain disease models. Yet, comprehensive automated observation methodologies remain scarce, and none have overbridged the issue of not always being able to track the full movement pattern independent of how the rodent twists and turns. This paper addresses this gap using a standardized recording setup with six inward-facing cameras. This setup not only enables a future complete 3D skeleton, overcoming the limitations of partial views and 2D analysis but also lays the foundation for incorporating detailed mapping of pose data to specific symptoms, potentially via the utilization of action recognition systems (Gharaee et al., 2017) combined with calcium imaging. This approach allows for continued research that could link observable behaviors with underlying neuronal activity, thereby facilitating the development of new therapeutic strategies.

1.1 Pose Estimation in Animal Studies

The introduction of DeepPose (Toshev and Szegedy, 2014) marked a leap in pose estimation techniques. It utilized deep convolutional neural networks (CNNs) to regress image pixels directly to spatial body joint locations, moving away from reliance on handcrafted features. Enhancements followed with methods combining CNNs and Markov Random Fields for modeling geometric and spatial constraints (Tompson et al., 2014), and incorporating temporal information to leverage movement continuity across frames (Pfister et al., 2015). Building on these advancements, position refinement models (Tompson et al., 2015) and the partitioning and labeling approach in DeepCut (Pishchulin et al., 2016) brought finer detail and improved detection capabilities. DeeperCut (Insafutdinov et al., 2016) leveraged the ResNet architecture to enhance accuracy and processing efficiency, introducing deep body part detectors that significantly improved precision.

DeepLabCut (Mathis et al., 2018) adapted DeeperCut's feature detector architecture for animal models, broadening pose estimation's applicability to non-invasive animal studies. This contributed to ethical research by minimizing stress and interference while maximizing analytical depth. Building upon the approach taken by DeepLabCut, LEAP (Pereira et al., 2019) focused on inference speed, while DeepPoseKit (Graving et al., 2019) introduced a model leveraging a stacked DenseNet architecture to improve speed and robustness.

For 3D pose estimation in animals, DeepFly3D (Günel et al., 2019) and Anipose (Karashchuk et al., 2021) advanced precision by introducing procedures for triangulating multiple cameras. LiftPose3D (Gosztolai et al., 2021) extended capabilities by adapting techniques designed for humans to animal models.

Recent contributions like Multi-animal DeepLabCut (Lauer et al., 2022) and SLEAP (Pereira et al., 2022) introduced multi-task architectures capable of identifying key points and tracking multiple animals simultaneously.

Despite significant progress, challenges remain due to the lack of annotated datasets for certain species and non-standardized recordings. Recent studies have proposed alternative approaches to address data scarcity. Biderman et al. introduced Lightning Pose, a semi-supervised model utilizing both labeled and unlabeled data, employing a multi-network architecture that leverages temporal and spatial contexts without extensive annotations (Biderman et al., 2023). Similarly, Li and Lee developed ScarceNet, which uses a pseudo-label-based approach, training a model with a small set of labeled images to generate pseudo-labels for unlabeled data (Li and Lee, 2023).

While promising, these techniques have not yet gained broad acceptance, and the field predominantly relies on supervised methods. This underscores the potential utility of the presented dataset.

1.2 Behavioral Analysis in Animals

The introduction of JAABA (Kabra et al., 2013) marked a shift towards using pose estimation for behavior classification in mice and flies. JAABA utilized pose trajectories to compute per-frame features, demonstrating the feasibility of behavior analysis using pose data.

MotionMapper (Berman et al., 2014) presented an unsupervised behavior classification pipeline for invertebrates, mapping behaviors to a 2D plane without prior labeling by segmenting, scaling, and aligning frames before applying PCA. This method identified distinct general behavioral modes but was limited to invertebrates.

MoSeq (Wiltschko et al., 2015) expanded unsupervised behavior analysis to vertebrates, specifically mice. By compressing video data and segmenting it into discrete behavioral ”syllables” using an autoregressive hidden Markov model (AR-HMM), MoSeq offered insights into the modular nature of animal behaviors, contributing to the understanding of complex behavioral patterns.

DeepBehavior (Arac et al., 2019) leveraged GoogLeNet and YOLO-v3 architectures to identify individual and social behaviors in mice. However, the focus remained on general social interactions rather than motoric analysis pertinent to degenerative diseases.

BehaveNet (Batty et al., 2019) introduced a probabilistic framework combining video compression with AR-HMM segmentation for unsupervised analysis. This allowed for simulating behavioral videos based on neural activity in mice, offering potential pathways for linking observable behaviors with underlying conditions, though not directly applied to motoric components.

SimBA (Nilsson et al., 2020) and MARS (Segalin et al., 2021) designed pose-based approaches for analyzing rodent social behavior, with SimBA extending tools to both mice and rats. While these studies provided valuable insights into social dynamics, they did not focus on the motoric components. DeepEthogram (Bohnslav et al., 2021) employed a supervised approach using optical flow estimated from video clips for behavior classification but similarly concentrated on social behaviors.

While pose estimation has enhanced our understanding of animal behavior, there is a gap in automated methods for motoric components of degenerative disease models. At the same time, there is no public rat dataset for detailed benchmarking at the desired granularity. Given the mobility of soft tissue and fur, introducing a standard could also minimize human bias and allow for synchronizing datasets between laboratories.

2 RECORDING FRAME

The recording apparatus is custom-designed to support a six-camera system that captures all six sides of a cubic space. This configuration, with inward-facing cameras mounted on the frame, enables comprehensive capture of rodent behavior within the enclosure.

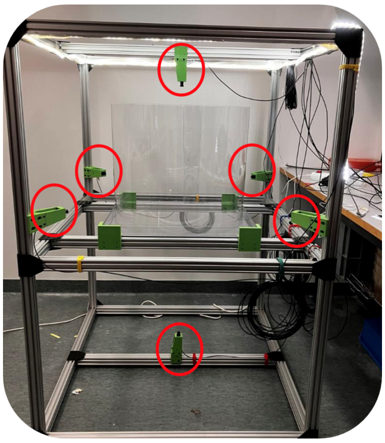

Figure 1: The recording frame with its six cameras and mounts (circled in red) facing the transparent arena.

We utilized NileCAM25 cameras for their capability to record high-definition (HD) video at 60Hz and support feed synchronization. These specifications ensure high-quality video data, which is critical for detailed pose annotation and precise movement tracking. At the system’s core is a Jetson AGX Orin, which processes camera input via a GMSL2 deserializer board. This setup streamlines data capture and synchronizes feeds from all six cameras, ensuring temporal alignment of footage. Synchronization is essential for enabling future 3D tracking of movement patterns. This approach ensures that pose predictions are conducted simultaneously across all perspectives, avoiding introducing spatial deviations in identified positions.

2.1 Video Recording, Frame Extraction, and Dataset Construction

Each of the 22 recorded sessions comprised six video streams corresponding to the six cameras, resulting in 132 MKV video files from the angles above/side/below.

Frame extraction was conducted post-recording using K-means to select the most diverse frames from each video. By setting k=50, we partitioned the video data into 50 clusters per video, extracting the centroid frame of each cluster. This approach maximizes the variety of poses and activities within the dataset, reducing bias toward any specific behavior or arena region where the rat might spend significant time.

From this process, 6.600 frames were extracted and further annotated with 35 labels representing various anatomical positions of the skeleton. The dataset was then partitioned into training, validation, and test sets with an 80-10-10 split. It was constructed solely using healthy control animals.

2.2 Data Augmentation

The employed augmentation methods are derived from the ‘packaging’ of DeepLabCut and adjusted to fit the project’s setup (Mathis et al., 2018). The following Table (1) and section outline the specific augmentation methods in each augmentation package and motivate their inclusion.

Table 1: Affiliation of augmentation methods for each investigated package (tensorpack, imgaug and default).

Augmentation	tensorpack	imgaug	default
Mirroring	✓		
Rotation	✓	✓	
Scaling		✓	✓
Motion Blur	✓		
Gaussian Noise	✓	✓	
Gaussian Blur		✓	
Elastic Transformation	✓		
Grayscale	✓		
Contrast Adjustments	✓	✓	
Filters	✓		
Crop	✓	✓	
Pad	✓		
Brightness		✓	
Covering	✓		
Saturation		✓	
Resize	✓		

Spatial transformations include mirroring, which creates horizontal flips of images to help models recognize and track poses regardless of the animal’s orientation, preventing bias toward the direction in which the animal most frequently appears. Rotation introduces angular perspectives, replicating the natural variance observed when an animal moves in three-dimensional space, ensuring accuracy across different capture angles. Covering (adding dropout regions) mimics occlusion events when a body part is temporarily hidden, training the model to infer poses and maintain tracking despite visual obstructions. Elastic transformations introduce image distortion, challenging the model to recognize anatomical features even when distorted by movement during rapid activities. Crop and pad operations introduce boundary variability, teaching the model to handle cases where the rat is not entirely within the frame or is positioned toward the edges. Resize ensures accurate pose estimation across different resolutions and scales, which is essential for processing data from various sources or camera types. Scaling allows for random resizes within a specified range, maintaining the aspect ratio but altering spatial dimensions, mimicking zoom effects during recording.

Blur or noise adjustments involve motion blur, simulating the effect of rapid movement or lower frame rates. This allows the models to regard poses even when image clarity is compromised. Gaussian blur provides controlled blur, approximating loss of sharpness, e.g., mimicking an out-of-focus rat moving fast. Adding Gaussian noise conditions the model to disregard random noise in the image, focusing on critical features, simulating sensor noise in low light.

Color and light adaptations encompass contrast adjustments, addressing scenarios where lighting alters the rat’s appearance. Brightness and saturation adjustments handle changes in visual perception due to environmental illumination and camera exposure. Converting images to grayscale forces the model to rely less on color and more on structural information, enhancing performance under varying conditions.

Filter effects simulate camera lens imperfections or environmental factors that might affect quality.

2.3 Post-Processing and Evaluation

The post-processing consists of three main steps: confidence filtering, moving Z-score outlier detection, and spline interpolation for filling in missing values.

The initial step involves applying a confidence filter to the raw pose estimation data. This threshold was set to 0.6 to adjust for a balance between minimizing false positives and the risk of losing true positives.

A moving Z-score is applied to identify and remove outliers following confidence filtering.

The interpolation of missing values resulting from the removal of low-confidence detections and outliers follows thereafter. Cubic spline interpolation is employed since it provides smooth, continuous curves that naturally fit the movements observed in the data. This method interpolates missing data points by fitting a series of cubic polynomials between known data points, ensuring that the first and second derivatives of the interpolated curves are continuous across the dataset. This is calculated as:

$$S(x) = a_n x^3 + b_n x^2 + c_n x + d_n, \quad \text{for } x_n \leq x \leq x_{n+1} \tag{1}$$

where $S(x)$ represents the spline function, and a_n, b_n, c_n, and d_n are the coefficients of the cubic polynomial between known points x_n and x_{n+1}.

The evaluation of the model’s performance relies on Root Mean Squared Error (RMSE), as expressed in equation 2. The metric is based on the Euclidean distance, in pixels, between the predicted positions and the corresponding ground-truth positions. This is calculated as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(gt_i - pred_i)^2} \tag{2}$$

where gt_i represents the ground truth position while $pred_i$ depicts the prediction.

3 NETWORK COMPARISON

The Resnet 50 and 152 models demonstrated rapid initial learning and good generalization when default and imgaug augmentation packages were applied (Figure 2: A, B, D and E). This suggests that these configurations can effectively capture and generalize from the learned patterns. When applying the tensorpack augmentation, both architectures portray higher losses and significant volatility (Figure 2: C and F), which implies that the augmentations may destabilize the learning process.

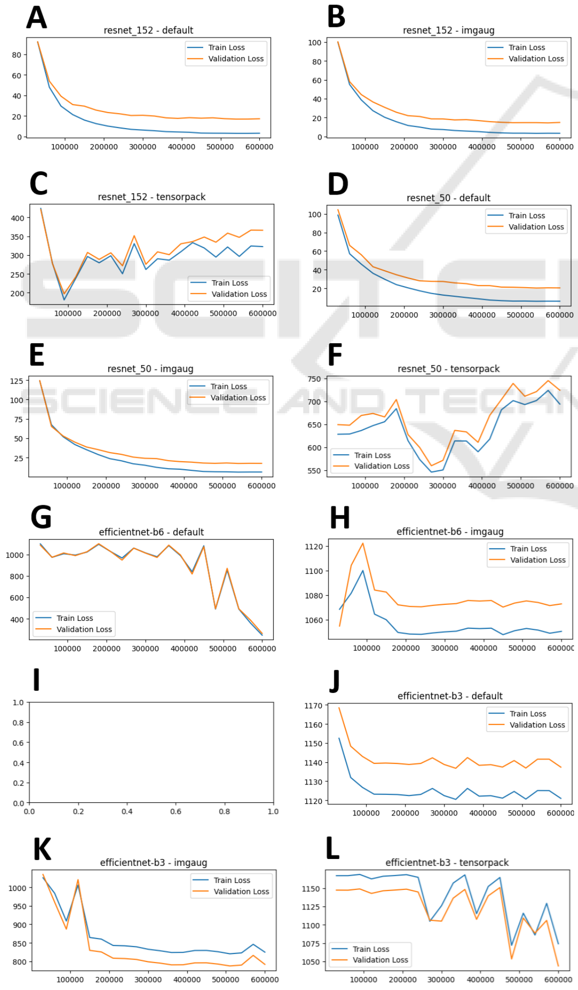

Figure 2: Network comparison, train- and validation loss for all network configurations A-L (architecture and augmentation method). Note that the Efficentnet B6 tensorpack (I) configuration did fail due to GPU memory constraints during training (Nvidia A100 node).

Among the various configurations, the Resnet 152 models with both default and imgaug augmentations emerge as the top performers (Table 2). Specifically, the Resnet 152 with default augmentation shows the lowest test error (16.81 px before the probability cutoff and 10.46 px after), making it the best-performing model expressed in RMSE. This is closely followed by the Resnet 152 with imgaug augmentation, demonstrating a performance of 14.31 px before and 10.48 px after the probability cutoff.

When examining configurations for the Efficientnet-B3 and B6 models, regardless of the augmentation used, they present a significantly higher test error than the leading Resnet models. This suggests challenges in the network architecture's ability to address the task and/or issues stemming from the augmentation methods. At the same time, the augmentation methods have contributed to low pixel error for the top-performing architectures, thus implying that the issue may lie on the architectural side.

This does not rule out that, given longer training sessions or further tuning, models like the Efficennet-B6 default could prove valid alternatives. With the obtained results, it is clear that the Resnet 152 default is the configuration to proceed with.

The tensorpack augmentation package includes a comprehensive set of transformations that significantly increase the complexity of the training data. These transformations are introduced to simulate real-world variances, such as changes in lighting, occlusions, and motion that a model might encounter in different implementation settings. The controlled environment represents a context where such variability does not often occur. This could explain why model configurations implementing tensorpack portray such issues. This indicates that even if tensorpack would be best suited for a more dynamic environment, the controlled experimental setup points to that lesser augmentation brings better model performance.

Although nominally small at $\sim 1.4\%$ pixels, the difference between the Resnet 50 and Resnet 152 default models represents an error increase of $\sim 12\%$. This provides a clear argument favoring the adoption of a deeper architecture.

The computational overhead is an aspect that should be considered in preclinical environments with limited computational resources or time constraints. The inference speed for processing six 5-minute videos via an Nvidia RTX 4080 for the Resnet 50 and Resnet 152 default models is shown in Table 3. From the results, an $\sim 63\%$ increase in processing time can be inferred when employing the deeper network.

Observational periods in studies of particular dis-

Table 2: Performance comparison of network configurations (model and augmentation method). Error is expressed in RMSE.

Model	Aug. method	Model best train iter.	% train set	Shuffle no.	Train err. (px)	Test err. (px)	p-cutoff	Train err. w. p-cutoff	Test err. w. p-cutoff
efficientnet-b3	default	330000	80	1	1120.47	1136.77	0.6	-	-
efficientnet-b3	imgaug	510000	80	1	820.38	787.55	0.6	-	-
efficientnet-b3	tensorpack	600000	80	1	1074.06	1043.82	0.6	-	-
efficientnet-b6	default	600000	80	1	246.13	262.68	0.6	452.19	333.3
efficientnet-b6	imgaug	30000	80	1	1068.39	1054.71	0.6	-	-
resnet 152	default	540000	80	1	2.93	16.81	0.6	2.75	10.46
resnet 152	imgaug	570000	80	1	3.33	14.31	0.6	3.15	10.48
resnet 152	tensorpack	90000	80	1	180.85	196.96	0.6	30.84	37.19
resnet 50	default	540000	80	1	5.99	20.25	0.6	4.33	11.85
resnet 50	imgaug	540000	80	1	6.25	17.43	0.6	4.29	12.01
resnet 50	tensorpack	270000	80	1	545.65	559.52	0.6	286.41	292.39

Table 3: Inference speed comparison for the two different Resnet depths investigated.

Model	Rec.	FPS	Dur. (s)	Inf. Speed (s)	Ratio
Resnet 152 default	6	30	1794	4602	1:2.6
Resnet 50 default	6	30	1794	2816	1:1.6

eases often extend to multiple hours per animal and require large quantities of animals to achieve significant results. The utilized recording frame adds a factor of 6 (cameras), thus introducing another layer of complexity that should be considered.

3.1 Single or Multiple Networks

The exploration of whether using multiple networks, each dedicated to a specific camera angle, would improve performance over a single network trained across all angles. For the dedicated ResNet 152 default networks (above, side, below), the training and validation loss curves showed rapid decreases followed by stabilization, indicating effective learning without overfitting (Figure 3).

From Table 4, it is clear that the multi-network approach results in lower test errors compared to the single-network.

While theoretically, free-moving rats could expose all body parts to any camera, the diversity captured using the frame extraction method (K-means) may not encompass all possible poses for each angle. This raises the question of whether disease-induced motor symptoms, like seizures, might affect the model's accuracy if the angle-specific networks have not been trained on such variations. Each disease model may require additional data to capture these behaviors. The multi-network approach could serve as a transfer basis, but this consideration applies regardless of whether single or multiple networks are used.

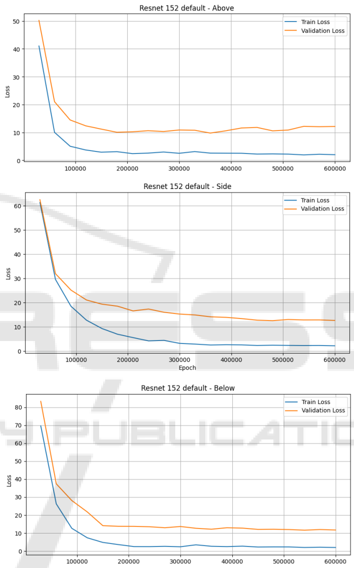

Figure 3: Resnet 152 default train- and validation loss seen for the dedicated networks (multi-network approach). From the top: above, side, and below.

3.2 Limitations

The findings are based on a controlled experimental environment designed for preclinical behavioral studies, which may limit the generalizability to more dynamic settings where other model configurations might perform better. Individuals performed data annotation, introducing potential biases in label placement and data quality assessments, though efforts were made to mitigate these issues.

The investigation was limited to selected network architectures, depths, and augmentation methods chosen for their relevance and proven performance. Al-

Table 4: Performance for the dedicated (angle-specific) models versus the single (one covering all angles) network. Error is expressed in RSME.

Model	Approach	Angle	Train err. (px)	Test err (px)	Train err. w/ p-cutoff	Test err. w/ p-cutoff
Resnet 152 default	Angle-Specific	Above	2.57	9.77	2.57	8.74
Resnet 152 default	Angle-Specific	Below	2.03	11.64	2.02	8.78
Resnet 152 default	Angle-Specific	Side	2.47	12.6	2.39	9.72
Resnet 152 default	Single	All	2.93	16.81	2.75	10.46

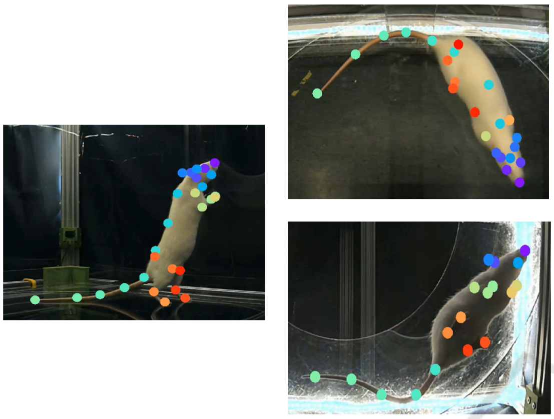

Figure 4: Illustration of pose detection from various camera angles using the dedicated networks. The detected poses align with the designated labels of the defined rat skeleton.

ternative configurations might yield better results. Consistent network parameter tuning was maintained across configurations to ensure comparability, but this may have restricted optimal tuning for each model.

4 CONCLUSIONS

This paper has explored the application of different deep-learning architectures, depths, and augmentation techniques for pose estimation in rodents, given the presented 6-camera recording frame. The focus has been on ResNet 50, ResNet 152, EfficientNet-B3, and EfficientNet-B6. In conclusion, ResNet 152, with default augmentation, is the most effective choice given the controlled experimental setup utilized.

The study also examined the efficacy of employing a single network trained across all six camera angles versus multiple networks, each dedicated to a specific camera angle. The findings favored the multi-perspective approach utilizing the Resnet 152 default configuration.

The results show that the configuration can construct coherent movement patterns in 2D from the detections, which further lays a foundational step towards achieving tracking in 3D. This enables a more granular analysis of the motoric components of degenerative diseases and opens up the possibility of extracting behavioral syllabus and potentially mapping the same to neural activity in the future.

4.1 Future Work

Continued work could explore improving the Efficientnet tuning and investigate how they handle different camera angles.

Future directions should assess how changes in network depth affect processing times, provide a more detailed mapping of the preclinical areas that require vast processing, and clarify the extent.

The approach could serve as a base for transfer learning, thus allowing adaptations for other anatomically similar species, such as mice. Therefore, another direction is to investigate how well the solution generalizes to mice and determine the required additional data.

A natural continuation of the project is to utilize the 2D data to construct a 3D representation of the rodents. This will allow for behavioral analysis, which, together with calcium imaging of brain activity, could allow further research into the neural activity associated with degenerative diseases.

ACKNOWLEDGEMENTS

We gratefully acknowledge the contributions of numerous students who were involved in data labeling, frame mounting, and similar tasks.

REFERENCES

Arac, A., Zhao, P., Dobkin, B. H., Carmichael, S. T., and Golshani, P. (2019). Deepbehavior: A deep learning toolbox for automated analysis of animal and human behavior imaging data. *Frontiers in systems neuroscience*, 13:20.

Batty, E., Whiteway, M., Saxena, S., Biderman, D., Abe, T., Musall, S., Gillis, W., Markowitz, J., Churchland, A., Cunningham, J. P., et al. (2019). Behavenet: nonlinear embedding and bayesian neural decoding of behavioral videos. *Advances in Neural Information Processing Systems*, 32.

Berman, G. J., Choi, D. M., Bialek, W., and Shaevitz, J. W. (2014). Mapping the stereotyped behaviour of freely moving fruit flies. *Journal of The Royal Society Interface*, 11(99):20140672.

Biderman, D., Whiteway, M. R., Hurwitz, C., Greenspan, N., Lee, R. S., Vishnubhotla, A., Warren, R., Pedraja, F., Noone, D., Schartner, M., et al. (2023). Lightning pose: improved animal pose estimation via semi-supervised learning, bayesian ensembling, and cloud-native open-source tools. *BioRXiv*.

Bohnslav, J. P., Wimalasena, N. K., Clausing, K. J., Dai, Y. Y., Yarmolinsky, D. A., Cruz, T., Kashlan, A. D., Chiappe, M. E., Orefice, L. L., Woolf, C. J., et al. (2021). Deepethogram, a machine learning pipeline for supervised behavior classification from raw pixels. *Elife*, 10:e63377.

Gharaee, Z., Gärdenfors, P., and Johnsson, M. (2017). Online recognition of actions involving objects. *Biologically inspired cognitive architectures*, 22:10–19.

Gosztolai, A., Günel, S., Lobato-Ríos, V., Pietro Abrate, M., Morales, D., Rhodin, H., Fua, P., and Ramdya, P. (2021). Liftpose3d, a deep learning-based approach for transforming two-dimensional to three-dimensional poses in laboratory animals. *Nature methods*, 18(8):975–981.

Graving, J. M., Chae, D., Naik, H., Li, L., Koger, B., Costelloe, B. R., and Couzin, I. D. (2019). Deepposekit, a software toolkit for fast and robust animal pose estimation using deep learning. *Elife*, 8:e47994.

Günel, S., Rhodin, H., Morales, D., Campagnolo, J., Ramdya, P., and Fua, P. (2019). Deepfly3d, a deep learning-based approach for 3d limb and appendage tracking in tethered, adult drosophila. *Elife*, 8:e48571.

Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., and Schiele, B. (2016). Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 34–50. Springer.

Kabra, M., Robie, A. A., Rivera-Alba, M., Branson, S., and Branson, K. (2013). Jaaba: interactive machine learning for automatic annotation of animal behavior. *Nature methods*, 10(1):64–67.

Karashchuk, P., Rupp, K. L., Dickinson, E. S., Walling-Bell, S., Sanders, E., Azim, E., Brunton, B. W., and Tuthill, J. C. (2021). Anipose: A toolkit for robust markerless 3d pose estimation. *Cell reports*, 36(13).

Lauer, J., Zhou, M., Ye, S., Menegas, W., Schneider, S., Nath, T., Rahman, M. M., Di Santo, V., Soberanes, D., Feng, G., et al. (2022). Multi-animal pose estimation, identification and tracking with deeplabcut. *Nature Methods*, 19(4):496–504.

Li, C. and Lee, G. H. (2023). Scarcenet: Animal pose estimation with scarce annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17174–17183.

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.

Nilsson, S. R., Goodwin, N. L., Choong, J. J., Hwang, S., Wright, H. R., Norville, Z. C., Tong, X., Lin, D., Bentzley, B. S., Eshel, N., et al. (2020). Simple behavioral analysis (simba)–an open source toolkit for computer classification of complex social behaviors in experimental animals. *BioRxiv*, pages 2020–04.

Pereira, T. D., Aldarondo, D. E., Willmore, L., Kislin, M., Wang, S. S.-H., Murthy, M., and Shaevitz, J. W. (2019). Fast animal pose estimation using deep neural networks. *Nature methods*, 16(1):117–125.

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., et al. (2022). Sleap: A deep learning system for multi-animal pose tracking. *Nature methods*, 19(4):486–495.

Pfister, T., Simonyan, K., Charles, J., and Zisserman, A. (2015). Deep convolutional neural networks for efficient pose estimation in gesture videos. In *Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part I 12*, pages 538–552. Springer.

Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P. V., and Schiele, B. (2016). Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4929–4937.

Segalin, C., Williams, J., Karigo, T., Hui, M., Zelikowsky, M., Sun, J. J., Perona, P., Anderson, D. J., and Kennedy, A. (2021). The mouse action recognition system (mars) software pipeline for automated analysis of social behaviors in mice. *Elife*, 10:e63720.

Tompson, J., Goroshin, R., Jain, A., LeCun, Y., and Bregler, C. (2015). Efficient object localization using convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 648–656.

Tompson, J. J., Jain, A., LeCun, Y., and Bregler, C. (2014). Joint training of a convolutional network and a graphical model for human pose estimation. *Advances in neural information processing systems*, 27.

Toshev, A. and Szegedy, C. (2014). Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660.

Wiltschko, A. B., Johnson, M. J., Iurilli, G., Peterson, R. E., Katon, J. M., Pashkovski, S. L., Abraira, V. E., Adams, R. P., and Datta, S. R. (2015). Mapping sub-second structure in mouse behavior. *Neuron*, 88(6):1121–1135.