nature neuroscience

SUPPLEMENTARY INFORMATION

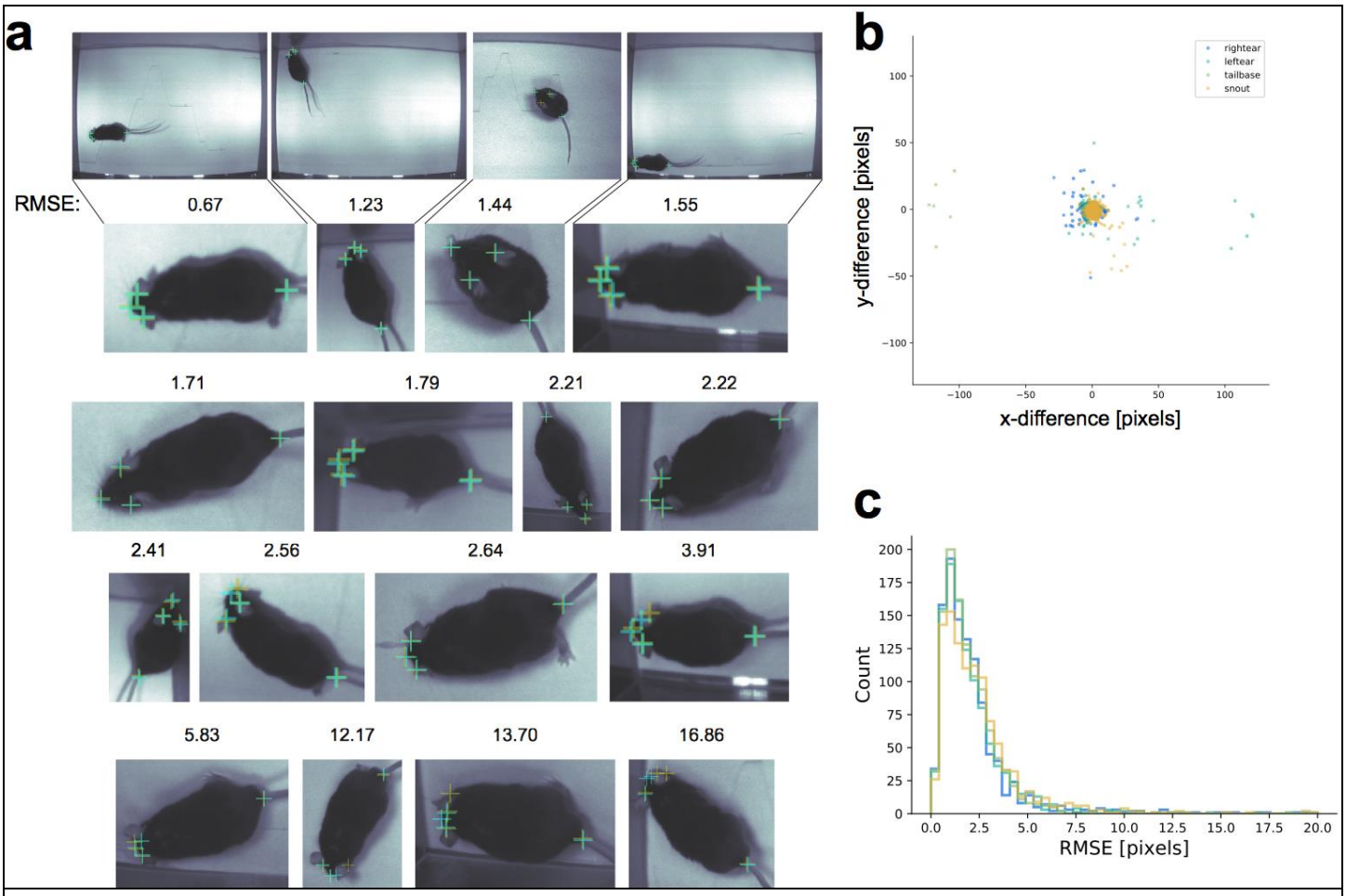
https://doi.org/10.1038/s41593-018-0209-y

In the format provided by the authors and unedited.

# DeepLabCut: markerless pose estimation of user-defined body parts with deep learning

**Alexander Mathis[1,2], Pranav Mamidanna[1], Kevin M. Cury[3], Taiga Abe[3], Venkatesh N. Murthy[2], Mackenzie Weygandt Mathis[1,4,8*] and Matthias Bethge[1,5,6,7,8]**

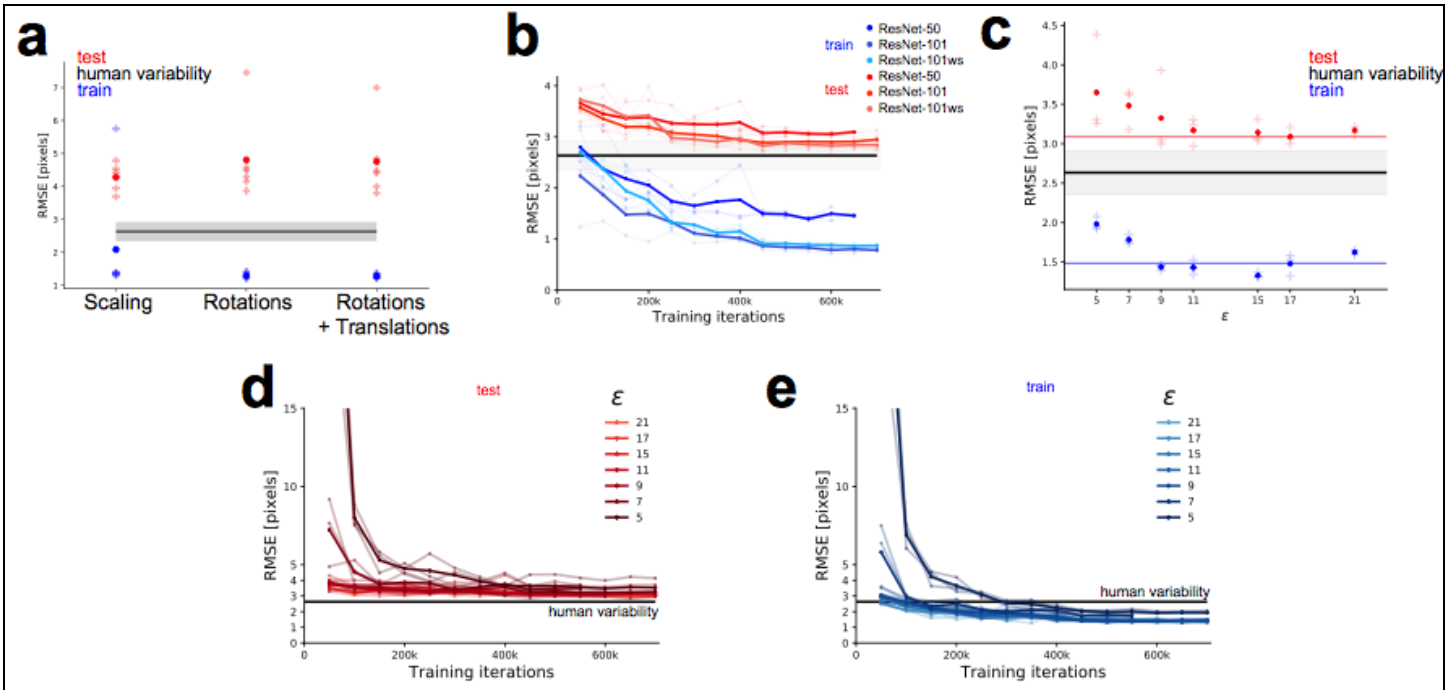
[1]Institute for Theoretical Physics and Werner Reichardt Centre for Integrative Neuroscience, Eberhard Karls Universität Tübingen, Tübingen, Germany. [2]Department of Molecular & Cellular Biology and Center for Brain Science, Harvard University, Cambridge, MA, USA. [3]Department of Neuroscience and the Mortimer B. Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY, USA. [4]The Rowland Institute at Harvard, Harvard University, Cambridge, MA, USA. [5]Max Planck Institute for Biological Cybernetics, Tübingen, Germany. [6]Bernstein Center for Computational Neuroscience, Tübingen, Germany. [7]Center for Neuroscience and Artificial Intelligence, Baylor College of Medicine, Houston, TX, USA. [8]These authors jointly directed this work: Mackenzie Weygandt Mathis, Matthias Bethge. *e-mail: mackenzie@post.harvard.edu




**Supplementary Figure 1**

**Illustration of labeled data set.**

**a**: Example images with human applied labels (first and second trial by the same labeler colored in yellow and cyan) illustrating the variability. Top row shows full frames that were labeled and illustrate typical examples of the 1,080 random frames, which comprise the data set. Rows below are cropped for visibility of the labels and indicate the average RMSE in pixels across all body parts above the image. The scorer was highly accurate, as illustrated by b. **b:** x-axis and y-axis difference in pixels between the first - second trial. Only a few labels strongly deviate between the two trials. Most errors are smaller than 5 pixels as can be seen in the histogram of trial-by-trial labeling errors (cropped at 20 pixels), **c**).




**Supplementary Figure 2**

**Cross-validating model parameters and testing deeper networks.**

**a:** Average training and test error for the same 6 splits with 10% training set size as in Figure 2f with standard augmentation (i.e. scaling), augmentation by rotations as well as rotations and translations (see Methods). Although with augmentation there are 8 and 24 times as many training samples (rotations and rotations + translations, respectively), the training and test errors remained comparable. **b:** Average training and test error for the same 3 splits with 50% training set size for three different architectures: ResNet-50, ResNet-101 as well as ResNet-101ws, where part loss layers are added to conv4 bank[29]. For these networks the training error is strongly reduced, and the test performance modestly improved, indicating that the deeper networks do not over-fit (but do not offer radical improvement). Averaged over 3 splits, individual simulation results shown in as faint lines. The deeper networks reach human level accuracy on test set. The data for ResNet-50 is also depicted in Figure 2d. **c:** Cross validating model parameters for ResNet-50 and 50%-training set fraction. We varied the distance variable ϵ, which determines the width of the score-map template during training around the ground-truth value with scale variable 100% (otherwise the scale ratio of the output layer was set to 80% relative to the input image size). Varying distance parameters only mildly improves the test performance (after 500k training steps). The average performance for scale 0.8 and ϵ=17 is indicated by horizontal lines (from Figure 2d). In particular, for smaller distance parameters the RMSE increases and learning proceeds much slower (c,d). **d-e:** Evolution of the training and test errors at various states of the network training for various distance variables ϵ corresponding to **c**.