Supplementary Figures and Movie Legends

Title: Fast animal pose estimation using deep neural networks

Authors:
Pereira, T. D.[1,#], Aldarondo, D. E.[1,#], Willmore, L.[1], Kislin, M.[1], Wang, S. S.-H.[1,2], Murthy, M.[1,2*], and Shaevitz, J. W.[1,3,4*]

1 Princeton Neuroscience Institute, Princeton University

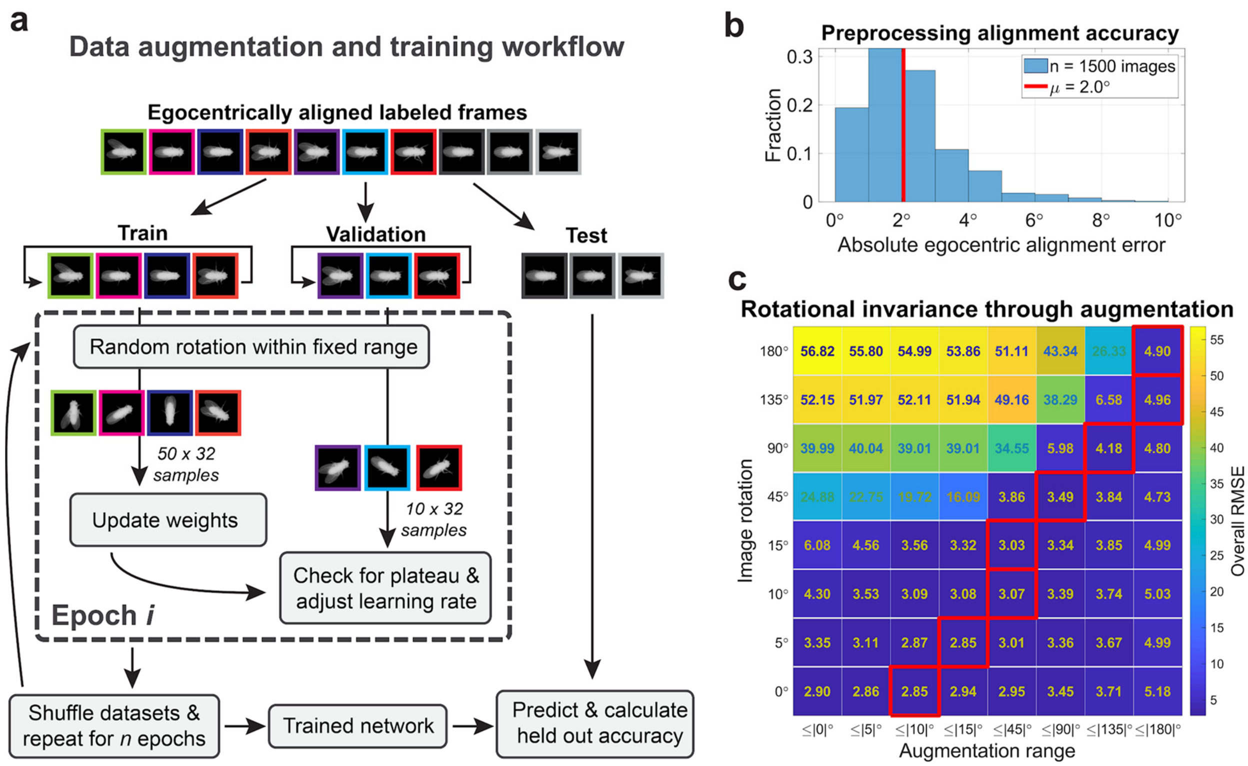
2 Department of Molecular Biology, Princeton University

3 Lewis-Sigler Institute for Integrative Genomics, Princeton University

4 Department of Physics, Princeton University

#equal authors

*lead contacts and co-corresponding authors: Mala Murthy (mmurthy@princeton.edu) and Joshua W. Shaevitz (shaevitz@princeton.edu)

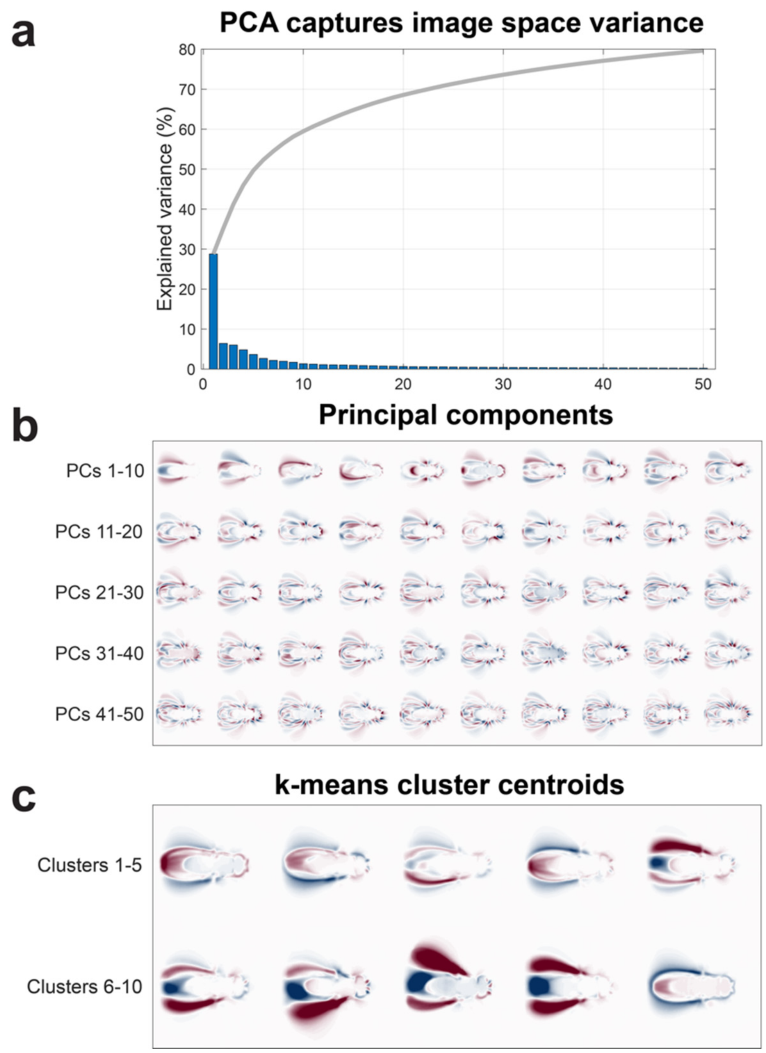


Supplementary Figure 1: Rotational invariance is learned at the cost of prediction accuracy

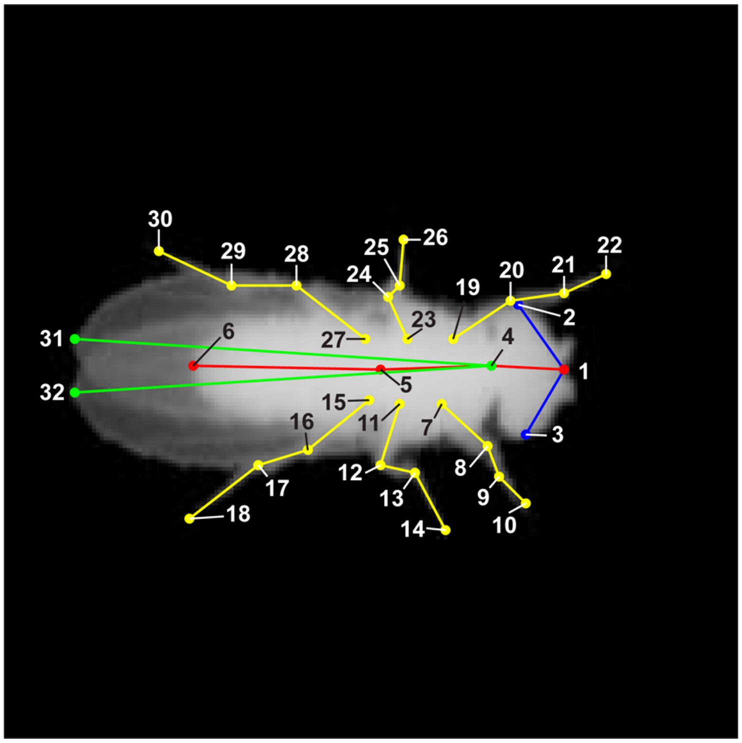
(a) The augmentation procedure consists of random rotations about the center of egocentrically aligned labeled frames. Labeled frames are split into training, validation and test sets. Colors are used to indicate unique images. Only training and validation sets are augmented and used for training.. During training, images are drawn sequentially from the training and validation sets to form batches of 32 images, cycling back to the beginning if there are less images than required, and then rotated randomly within a range of angles; confidence maps are rotated accordingly (not shown). After each epoch, the ordering of the datasets are shuffled so as to create new combinations of batches. The test set images are not augmented before computing accuracy metrics reported throughout.

(b) Egocentric alignment accuracy of the preprocessing algorithm from [1] when compared to manual labels of head/thorax. The error is the absolute deviation of the angle formed between the thorax and head from the horizontal centerline in the image. The mean of 2.0° indicates that there is little alignment error to which the network has to learn robustness.

(c) The accuracy measured as the RMSE of position estimates when evaluated on data artificially rotated at a fixed angle (rows) with networks trained on data augmented by rotations between a range of angles (columns). Red boxes denote the best accuracy for each data angle, denoting that optimal performance is achieved when the network is trained on augmented images with the minimally inclusive range of angles. Top accuracy decreases relative to the degree of rotational invariance the network must learn.



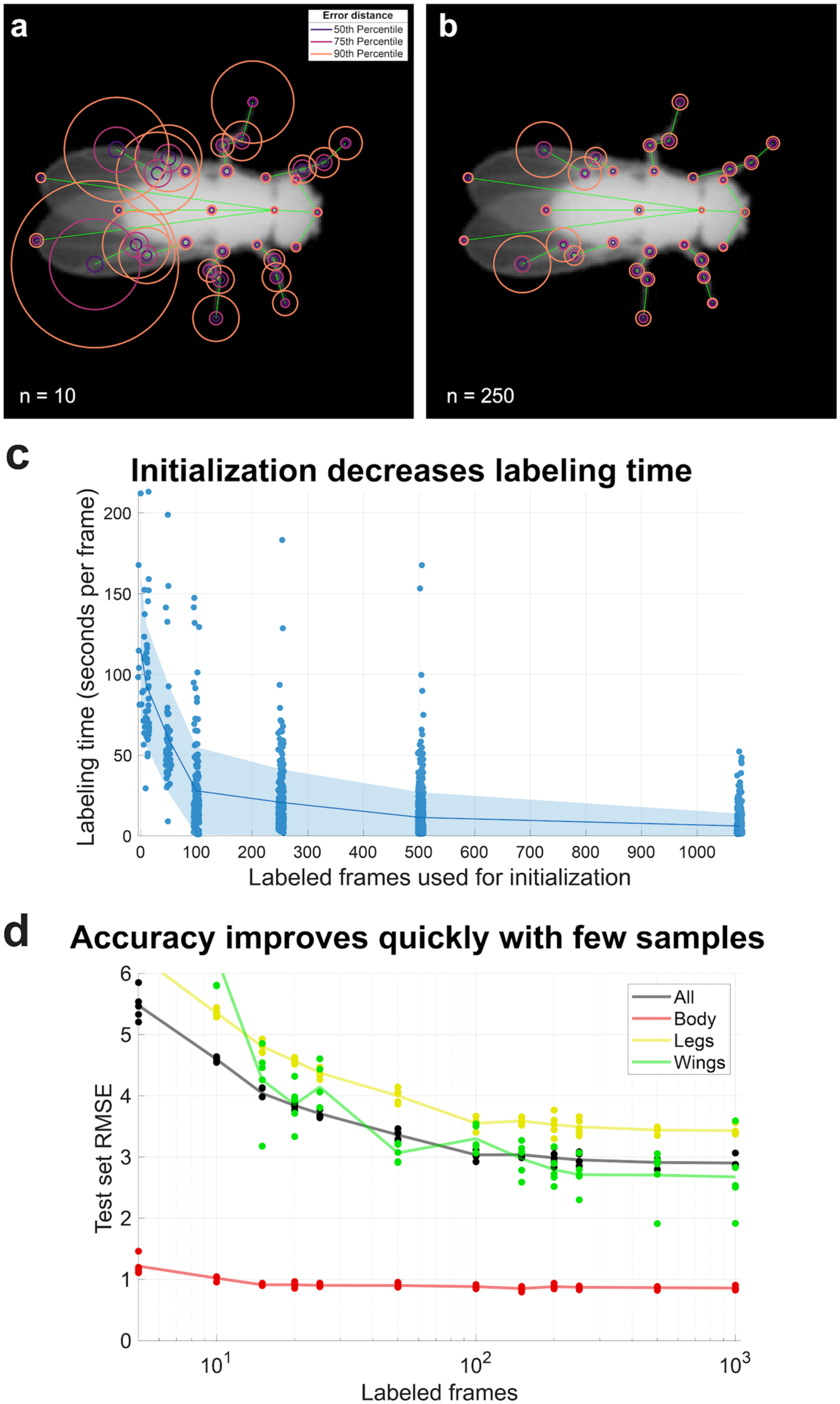
Supplementary Figure 2: Cluster sampling to promote pose diversity in labeling dataset
(a) Principal component analysis (PCA) of unlabeled images captures the majority of the variance in the data within 50 components. The cumulative variance explained (line) suggests that using PCA for dimensionality reduction does not sacrifice substantial information within the images.
(b) Top PCA eigenmodes visualized as coefficient images. Red and blue shading denote positive and negative coefficients at each pixel. Areas of similar colors indicate correlated pixel intensities within a given mode. After mean subtraction, each image in the initially sampled dataset is projected onto the first 50 eigenmodes.
(c) Cluster centroids identified by k-means after PCA. Red and blue shading denote pixels with higher or lower intensity than the overall mean. Cluster centroids illustrate the diversity of poses that are detected in image space by this sampling method. Samples are then drawn evenly from each cluster to select representative images for labeling with the GUI.



(1) Tip of head (between antennae)
(2) Left eye
(3) Right eye
(4) Neck / connects head and thorax
(5) Mesothoracic phragma (connects thorax and abdomen)
(6) Tip of abdomen
(7) Right foreleg / thorax-coxa
(8) Right foreleg / coxa-femur
(9) Right foreleg / femur-tibia
(10) Right foreleg / tarsus tip
(11) Right midleg / thorax-coxa
(12) Right midleg / coxa-femur
(13) Right midleg / femur-tibia
(14) Right midleg / tarsus tip
(15) Right hindleg / thorax-coxa
(16) Right hindleg / coxa-femur
(17) Right hindleg / femur-tibia
(18) Right hindleg / tarsus tip
(19) Left foreleg / thorax-coxa
(20) Left foreleg / coxa-femur
(21) Left foreleg / femur-tibia
(22) Left foreleg / tarsus tip
(23) Left midleg / thorax-coxa
(24) Left midleg / coxa-femur
(25) Left midleg / femur-tibia
(26) Left midleg / tarsus tip
(27) Left hindleg / thorax-coxa
(28) Left hindleg / coxa-femur
(29) Left hindleg / femur-tibia
(30) Left hindleg / tarsus tip
(31) Left wing tip
(32) Right wing tip

Supplementary Figure 3: User-defined skeleton

We selected 32 points to cover the body parts of the fly; these parts were chosen to approximately match the set of visible joints and interest points in the anatomy of the animal.

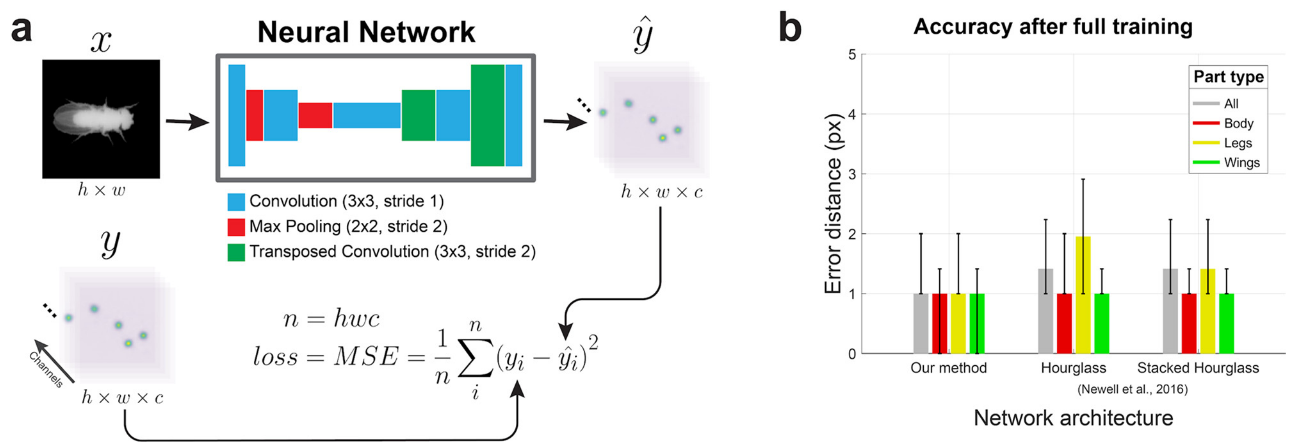


Supplementary Figure 4: Estimation accuracy improves with few samples

(a-b) Error distance distributions per body part when estimated with networks trained for 15 epochs on 10 (a) or 250 (b) labeled frames. The majority of estimates fall within few pixels of the ground truth, reducing the labeling procedure to simply correcting estimates.

(c) Time spent labeling each frame decreases with the quality of initialization. Line and shaded region correspond to mean and standard deviation. Starting frames require 115.4+-45.0 (mean+-s.d.) seconds to label, decreasing to 6.1±7.7 seconds after initializing with a network trained on 1000 labeled frames.

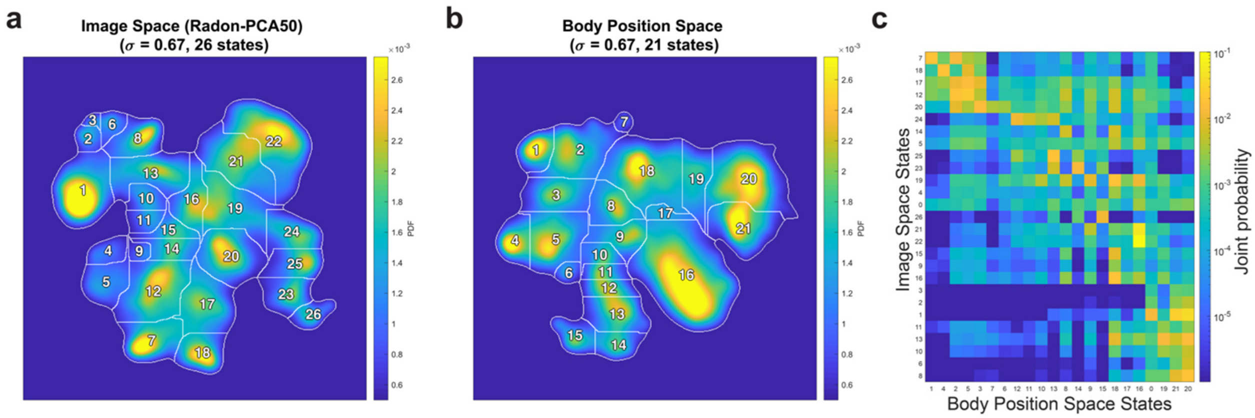
(d) Large accuracy improvements are observed with very few labeled samples, corresponding with the decrease in time required to fix initial labels on new frames. A plateau is observed at around 150-200 frames, with marginal improvements with additional labeling. Circles denote the test set RMSE for one replicate of fast training (15 epochs) at each dataset size, lines denote mean of all replicates.



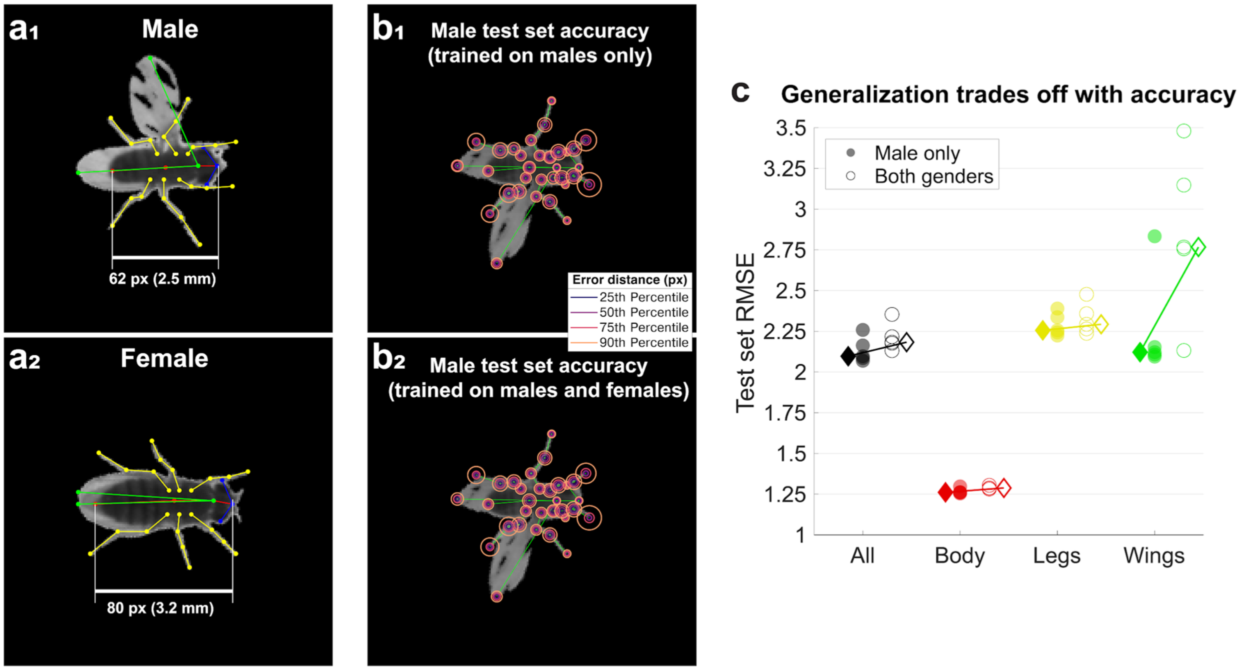
Supplementary Figure 5: Neural network architecture comparison

(a) Diagram of our neural network architecture. Raw images are provided as input into the network, which then computes a set of confidence maps of the same height and width as the input image (top row). The network consists of a set of convolutions, max pooling and transposed convolutions whose weights are learned during training (top middle). Estimated confidence maps are compared to ground truth maps generated from user labels using a mean squared error loss function, which is then minimized during training (bottom row).

(b) Accuracy comparison between architectures. We compared the accuracy of our architecture to the hourglass and stacked hourglass versions of the network described in[2]. The accuracy of our network is equivalent or better than those achieved when training with these reference architectures (over all body parts, $p < 1e-10$, Wilcoxon rank sum test, 1-tailed). Bar and error bars denote median and 25th and 75th percentiles.



Supplementary Figure 6: Comparison of behavioral space distributions generated from compressed images versus body part positions.
(a) Behavioral space distribution from 59 male flies calculated using the original MotionMapper pipeline (data and pipeline from [3]), including Radon-transform compression and PCA-based projection onto the first 50 principal components followed by a nonlinear embedding of the resultant spectrograms.
(b) Behavioral space distribution from 59 male flies (data and pipeline from [3]) calculated using spectrograms generated from tracked body part positions rather than PCA modes (see **Online Methods**). We note that this distribution has fewer peaks than that from (a) and a more symmetric topology (e.g in the top-left clusters, **Fig. 4c-g**).
(c) Joint probability distribution of the cluster labels from (a) and (b); sorted by row and column peaks. Many clusters identified using the pixel-based representation (rows) match up with those of the position-based representation (columns), but some are distributed into newly separated clusters.

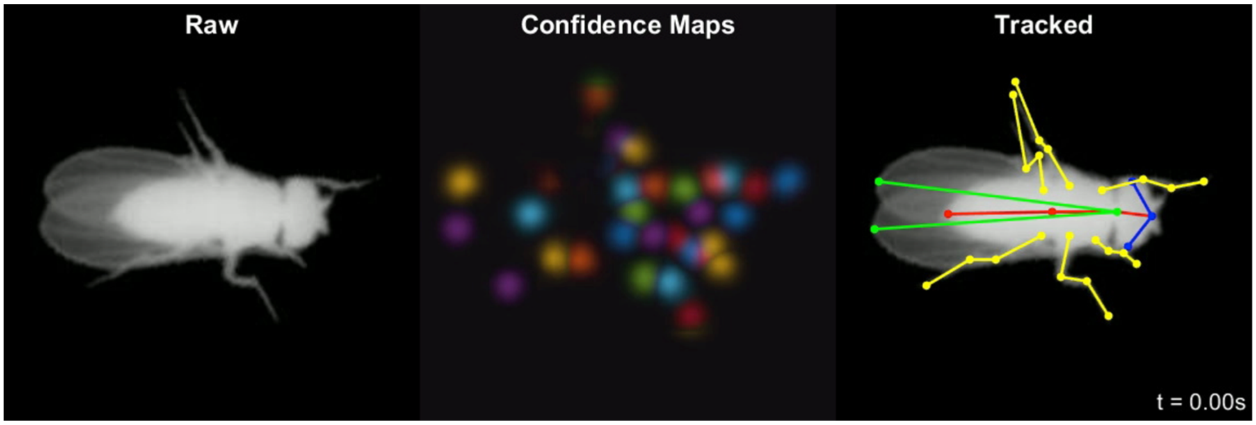


Supplementary Figure 7: Generalization to more diverse morphologies with a single network trades off with accuracy.

(a) Male and female flies differ in anatomical morphology, in part due to differences in their body length. The males (a_1) more often extend their wings as they are used to produce courtship song. The females (a_2) rarely extend their wings in this context, resulting in different requirements for pose estimation between the two genders, despite their overall similarity in morphology.

(b) Training on labeled images of just males (b_1) results in similar accuracy (on male test set images) to when training on both males and females (b_2). This suggests that there is little discernible difference (up to the 90th percentile) of having a network trained on two different types of body morphologies.

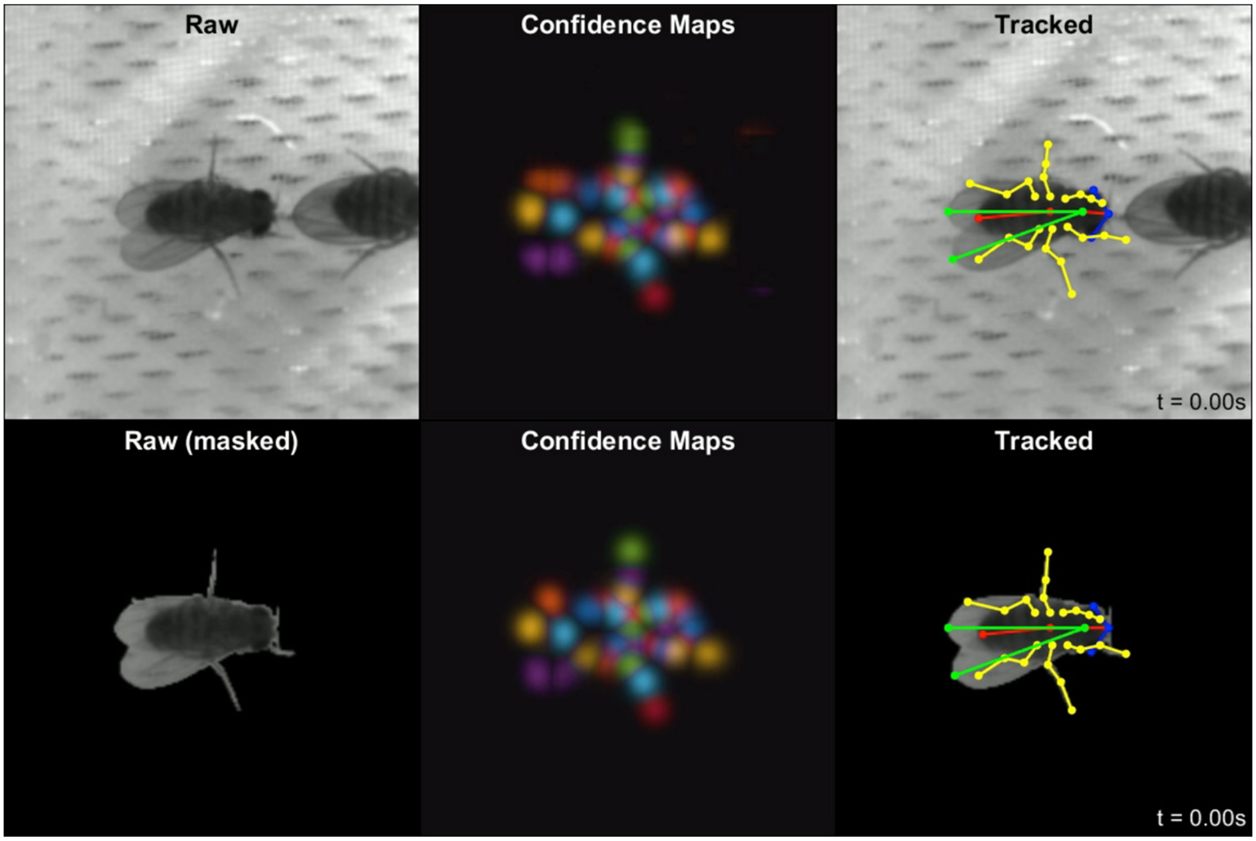
(c) Quantification of RMSE on the male test set shows that generalization to two different morphologies increases the error metric. Circles denote training replicates, diamonds denote median RMSE for all replicates, and filled and empty markers correspond to specialized versus generalized training respectively. Although the increased error rate is very small overall when generalizing, the greatest difference is observed in the body parts with greater difference in pose distributions (wings, green).



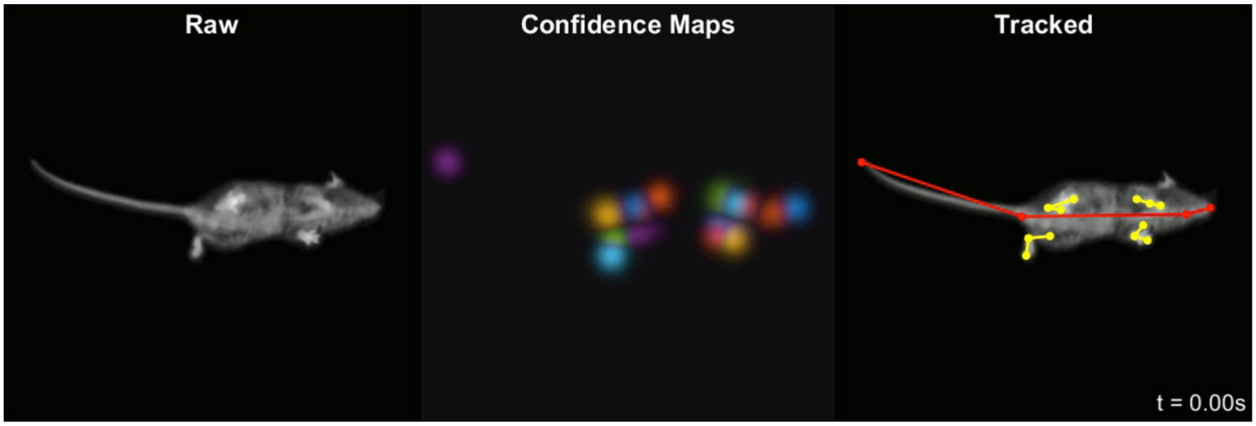
Supplementary Movie 1: Body part tracking is reliable over long periods without temporal constraints.
Raw images (left), max projection of all confidence maps (center), and tracked images (right) during a 20 second bout of free movement. Video playback at 0.2x realtime speed.

Supplementary Movie 2: Body part tracking during freely moving locomotion.
Raw images (left), max projection of all confidence maps (center), and tracked images (right) during a bout of locomotion. Video playback at 0.15x realtime speed. Video corresponds to Fig. 1d.

Supplementary Movie 3: Body part tracking during head grooming.
Raw images (left), max projection of all confidence maps (center), and tracked images (right) during a bout of head grooming. Video playback at 0.15x realtime speed. Video corresponds to Fig. 1e.



Supplementary Movie 4: Tracking joints robustly in images with heterogeneous background and noisy segmentation.
Raw images (left), max projection of all confidence maps (center), and tracked images (right) of a freely moving courting male fly. Rows correspond to results from a network trained on unmasked and masked images, respectively. Video playback at 0.2x realtime speed.



Supplementary Movie 5: Tracking joints in freely moving rodents.
Raw images (left), max projection of all confidence maps (center), and tracked images (right) of a freely moving mouse in an open field arena imaged from below through a clear acrylic floor. Video playback at 0.2x realtime speed. Tracking is reliable over time but degenerate when certain parts are occluded, such as when the animal rears.

References

1. Berman, G. J., Bialek, W. & Shaevitz, J. W. Predictability and hierarchy in Drosophila behavior. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 11943–11948 (2016).

2. Newell, A., Yang, K. & Deng, J. Stacked Hourglass Networks for Human Pose Estimation. in *Computer Vision – ECCV 2016* 483–499 (Springer International Publishing, 2016).

3. Berman, G. J., Choi, D. M., Bialek, W. & Shaevitz, J. W. Mapping the stereotyped behaviour of freely moving fruit flies. *J. R. Soc. Interface* **11,** (2014).