Animal pose estimation from video data with a hierarchical von Mises-Fisher-Gaussian model

Libby Zhang Stanford University
Timothy Dunn Duke University
Jesse Marshall Harvard University
Bence Ölveczky Harvard University
Scott Linderman Stanford University

Abstract

Animal pose estimation from video data is an important step in many biological studies, but current methods struggle in complex environments where occlusions are common and training data is scarce. Recent work has demonstrated improved accuracy with deep neural networks, but these methods often do not incorporate prior distributions that could improve localization. Here we present GIMBAL: a hierarchical von Mises-Fisher-Gaussian model that improves upon deep networks' estimates by leveraging spatiotemporal constraints. The spatial constraints come from the animal's skeleton, which induces a curved manifold of keypoint configurations. The temporal constraints come from the postural dynamics, which govern how angles between keypoints change over time. Importantly, the conditional conjugacy of the model permits simple and efficient Bayesian inference algorithms. We assess the model on a unique experimental dataset with video of a freely-behaving rodent from multiple viewpoints and ground-truth motion capture data for 20 keypoints. GIMBAL extends existing techniques, and in doing so offers more accurate estimates of keypoint positions, especially in challenging contexts.

1 Introduction

Studies of animal behavior are foundational to a wide range of disciplines, from psychology and neuroscience to drug discovery, ecology, and biomechanics (Dell et al., 2014; Krakauer et al., 2017; Brown and Bolivar, 2018; Musall et al., 2019). Despite its importance, obtaining precise, quantitative descriptions of animal behavior remains a challenge. Some techniques extract kinematic summary statistics from video, like the position and velocity of an animal's centroid (Jhuang et al., 2010; Gomez-Marin et al., 2012); others model the temporal dynamics of image features (Berman et al., 2014; Wiltschko et al., 2015). Advances in computer vision and human pose estimation have translated into new tools for automated, markerless, animal pose estimation (Mathis et al., 2018; Pereira et al., 2019; Graving et al., 2019; Wu et al., 2020).

While human pose estimation provides an important point of reference, studies of animal pose estimation in laboratory settings differ in important ways. Scientists have more control over the camera placement, calibration and visual environment and work with a relatively small number of similar individuals. However, scientific studies have lower tolerance for measurement error and a greater need for robustness than proof-of-principle studies in machine vision. Making reproducible inferences about animal behavior and simultaneously measured biological covariates requires 3D and temporally continuous measurements of pose, across diverse naturalistic behaviors. Furthermore, unlike in humans, large databases of annotated animal keypoints, especially in 3D, are only beginning to emerge (e.g. Bala et al., 2020).

In light of these constraints, effective tools for 3D markerless pose estimation in laboratory animals should possess high accuracy and spatiotemporal continuity, and be highly sample efficient with training data. Deep learning methods for pose estimation will help meet these goals, but the dual challenges of more stringent constraints and reduced availability of training data are likely to require statistical models that incorporate physical constraints of animal pose and leverage the stereotypy of animal behavior.

Recent work has made progress toward these goals. One line of work uses standard computer vision methods (Hartley and Zisserman, 2003) to produce 3D estimates from 2D keypoint estimates. For example, Nath et al. (2019) extended DeepLabCut—a popular 2D markerless tracking software package—to tri-

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

angulate multiple 2D markerless pose estimates, and Karashchuk et al. (2020) explored methods to make these estimates more robust. Other approaches incorporate 3D-awareness into their models, with varying degrees of model capacity and complexity. For example, Günel et al. (2019) used graphical models to enforce geometric consistency; Dunn et al. (2020) designed a volumetric convolutional neural network that enforces 3D geometric constraints; and Biggs et al. (2018) and Kearney et al. (2020) leveraged silhouette and 3D skinned-meshes to improve their model predictions. Kearney et al. (2020) incorporated a prior on 3D pose in the form of a hierarchical Gaussian process latent variable model, which learned joint rotations, heading, and locations. We develop a simpler and more explicit model of 3D pose that directly captures geometric constraints for animals with rigid skeletons.

We call this model GIMBAL: GeometrIc Manifolds for Body Articulation and Localization. GIMBAL uses recent advances in spherical manifold learning to capture the spatiotemporal constraints of body posture, and then leverages these constraints to accurately triangulate 3D keypoint locations from multiple 2D estimates. The central component of GIMBAL is a hierarchical von Mises-Fisher-Gaussian model, which lends itself to efficient Bayesian inference.

2 Background

Triangulating 3D location from 2D estimates Triangulation is a classic computer vision problem (Hartley and Zisserman, 2003). Consider a collection of videos from C calibrated cameras (fig. 1A). For each camera c we know the projective mapping $f_c : \mathbb{R}^3 \to \mathbb{R}^2$, which maps positions in 3D world-coordinates (in millimeters from a chosen origin) to locations on the camera image (in pixels from the image origin). The mapping amounts to an affine transform followed by a projective transformation, which we simplify as

$$f_c(x) = \tfrac{1}{w}(u, v)^\top \text{ where } (u, v, w)^\top = A_c x + b_c, \quad (1)$$

where A_c and b_c are known parameters of camera c. The affine transformation maps x into homogeneous coordinates of camera c; the nonlinear operation (dividing by w) converts the homogeneous coordinates into 2D positions in the image frame.

We will take a Bayesian approach to triangulation, assuming a prior distribution on the 3D locations and a noise model for the 2D estimates. Let K denote the number of keypoints, let $x_{t,k} \in \mathbb{R}^3$ denote the location at time t of keypoint k in world coordinates (in millimeters), and let $y_{t,k,c} \in \mathbb{R}^2$ denote the estimated location (in pixels) at time t of keypoint k in camera c. In general, let bold symbols denote sets of variables; e.g. $\boldsymbol{x} = \{\{x_{t,k}\}_{k=1}^K\}_{t=1}^T$ and $\boldsymbol{y} = \{\{y_{t,k,c}\}_{c=1}^C\}_{k=1}^K\}_{t=1}^T$. The posterior distribution on 3D locations is,

$$p(\boldsymbol{x} \mid \boldsymbol{y}) \propto p(\boldsymbol{x}) \prod_{t=1}^{T} \prod_{k=1}^{K} \prod_{c=1}^{C} p\big(y_{t,k,c} \mid f_c(x_{t,k})\big), \quad (2)$$

assuming the 2D estimates in each camera are conditionally independent given the 3D location.

This formulation allows for many different choices about the prior and likelihood. When both are assumed to be Gaussian, solving for the posterior mode amounts to solving a nonlinear least squares problem. However, outliers in the 2D estimates can have a large effect on the 3D inferences, so heavy-tailed noise models and other robust estimation techniques are recommended (Hartley and Zisserman, 2003).

Incorporating spatiotemporal constraints Another way to improve 3D estimation is to incorporate constraints and inductive biases into the prior distribution. For example, it is reasonable to assume that keypoints cannot move too far between consecutive frames. Such constraints can be encoded with a prior of the form,

$$p(\boldsymbol{x}) \propto \prod_{t=2}^{T} \prod_{k=1}^{K} \mathcal{N}(x_{t,k} \mid x_{t-1,k}, \eta_k^2 I), \quad (3)$$

where $\mathcal{N}$ denotes the Gaussian density and η_k^2 specifies the conditional variance of keypoint k.

Likewise, geometric constraints based on the physical distance between joints are common in pose estimation (Felzenszwalb and Huttenlocher, 2005; Yang and Ramanan, 2011; Amin et al., 2013; Burenius et al., 2013; Belagiannis et al., 2014; Pavlakos et al., 2017). These models, often called pictorial structures or deformable mixture of parts, penalize squared errors of estimated versus expected distances between keypoints. These penalties are equivalent to the prior,

$$p(\boldsymbol{x}) \propto \prod_{t=1}^{T} \prod_{(j,k)\in\mathcal{G}} \mathcal{N}\left(\|x_{t,k} - x_{t,j}\|_2 \mid \rho_{j,k}, \sigma_{j,k}^2\right), \quad (4)$$

where $\mathcal{G}$ is a undirected graph on the K keypoints, $\rho_{j,k}$ is the expected distance between keypoints j and k, and $\sigma_{j,k}^2$ is its variance.

Prior distributions like these can be combined and elaborated upon in various ways. We will construct a hierarchical generative model of 3D pose by extending recent methods for modeling curved manifolds, which naturally arise from geometric constraints.

The von Mises-Fisher-Gaussian distribution Mukhopadhyay et al. (2019) proposed an elegant approach to modeling data that lie near a curved manifold. Their method uses the von Mises-Fisher-Gaussian

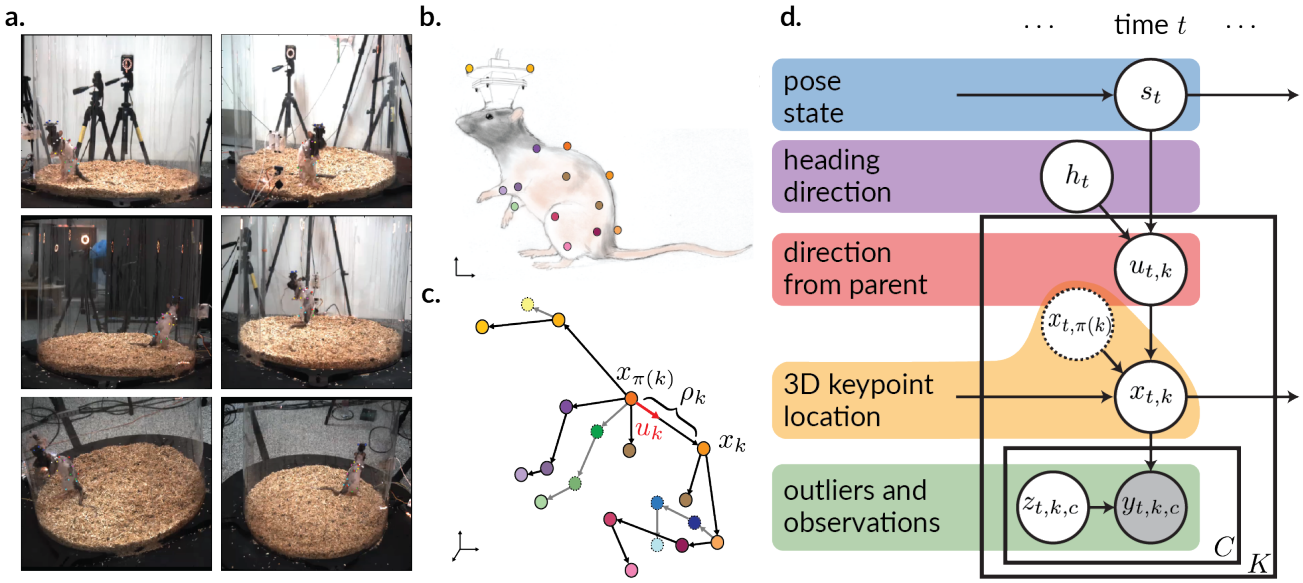

Figure 1: GIMBAL estimates 3D pose from multiple 2D videos with a hierarchical von Mises-Fisher-Gaussian model. **a.** The data consists of 2D keypoint estimates (colored dots) from multiple calibrated video streams. **b.** In each frame, a tracking algorithm estimates the 2D locations of visible keypoints. **c.** GIMBAL estimates the 3D keypoint locations x_k by incorporating geometric constraints in the form of a prior distribution that models directions u_k and distances ρ_k from one keypoint to another. **d.** The complete graphical model consists of a sequence of rotationally-invariant pose states, which, together with the heading direction, specify von Mises-Fisher distributions on directions between the K keypoints. The directions, together with past keypoint locations, parameterize a Gaussian model on the time series of 3D keypoint locations. The observed 2D locations are modeled as noisy projections of the 3D positions onto C calibrated cameras, and an outlier model affords robustness to misestimated 2D locations.

(vMFG) distribution, a compound distribution arising from a von Mises-Fisher mixture of Gaussians,

$$u \sim \text{vMF}(\nu, \kappa) \tag{5}$$

$$x \mid u \sim \mathcal{N}(\mu + \rho u, \sigma^2 I) \tag{6}$$

where $\text{vMF}(\nu, \kappa)$ denotes a von Mises-Fisher distribution with mean direction ν and concentration κ. The marginal distribution of x is a continuous mixture of Gaussians with means on the surface of a sphere centered at μ and radius ρ. Samples of x are thus concentrated around the surface of the sphere with variance set by σ^2 and direction governed by ν and κ.

Note that when the concentration κ is zero, the vMF reduces to a uniform distribution on the sphere. In that case, the marginal probability of x is only a function of the distance $\|x - \mu\|_2$, just as in eq. (4), so a special case of this model yields a distribution on distances.[1] The general formulation allows for directional priors as well, which GIMBAL will exploit.

Mukhopadhyay et al. (2019) showed that the von Mises-Fisher and Gaussian densities are conjugate in the sense that the conditional distribution of u given x is also a vMF, and the marginal probability of x has a closed-form expression in terms of normalizing constants of the vMF distribution. We build on these insights to develop a hierarchical von Mises-Fisher-Gaussian model for animal pose estimation and a simple algorithm for posterior inference.

3 Model

GIMBAL is a robust model for Bayesian triangulation of articulated body poses. The central component is a hierarchical von Mises-Fisher Gaussian model. We introduce models of postural dynamics, heading, and outliers in the 2D keypoints around this focal point.

Hierarchical von Mises-Fisher Gaussian model In many animal pose estimation tasks, keypoints of interest are connected to one another by rigid bones, which constrain the distances between them. We capture these dependencies with a tree-structured graph $\mathcal{G} = \{(\pi(k), k)\}_{k=2}^{K}$ where the keypoints are ordered such that keypoint 1 is the root node and each subsequent keypoint $k > 1$ has one parent $\pi(k) \in \{1, \ldots, k-1\}$. Roughly, the tree reflects the animal's skeleton (fig. 1b,c). We associate each edge $(\pi(k), k)$ with a length $\rho_k > 0$ specifying the average distance between those keypoints in 3D. Actual distances will vary since keypoints do not directly correspond to endpoints of rigid bones, but rather to points on the animal's skin. The von Mises-Fisher-Gaussian distribution is well-suited to modeling these skeletal

[1] In this case, however, the distance distribution is not Gaussian. See Section S1 in the supplementary materials.

constraints while allowing some flexibility in precise distances.

We place a hierarchical prior on the 3D keypoint positions that leverages the skeleton structure. The root node is modeled as a random walk, $x_{t,1} \sim \mathcal{N}(x_{t-1,1}, \eta_1^2 I)$, where η_1^2 controls the variance of movement between frames, as in (3) above. For keypoints $k > 1$, we combine temporal information from past locations and spatial information from parent locations to obtain the conditional distribution,

$$\begin{aligned} &p(x_{t,k} \mid x_{t-1,k}, x_{t,\pi(k)}, u_{t,k}) \\ &\quad \propto \mathcal{N}(x_{t,k} \mid x_{t-1,k}, \eta_k^2)\, \mathcal{N}(x_{t,k} \mid x_{t,\pi(k)} + \rho_k u_{t,k},\, \sigma_k^2) \\ &\quad = \mathcal{N}(\tilde{\mu}_{t,k}, \tilde{\sigma}_{t,k}^2), \end{aligned} \tag{7}$$

where

$$\begin{aligned} \tilde{\mu}_{t,k} &= \alpha_k \cdot x_{t-1,k} + (1 - \alpha_k) \cdot (x_{t,\pi(k)} + \rho_k u_{t,k}) \\ \alpha_k &= \eta_k^{-2} / (\eta_k^{-2} + \sigma_k^{-2}) \\ \tilde{\sigma}_k^2 &= 1 / (\eta_k^{-2} + \sigma_k^{-2}). \end{aligned}$$

The conditional mean is a convex combination of the past location and the offset from the parent node, and the weight of these two pieces of information is determined by their variances. Compare this model to the vMFG model above, and note that marginalizing over the direction vector under a uniform prior yields a conditional distribution proportional to a vMFG centered on the parent, tilted by a Gaussian centered on the preceding location. Hence, we call this a hierarchical von Mises-Fisher-Gaussian model.

Pose states and directional priors. Next, we specify a prior distribution on the direction vectors $\boldsymbol{u}$ in terms of a sequence of discrete *pose states* $\boldsymbol{s} = s_1, \ldots, s_T$, where $s_t \in \{1, \ldots, S\}$. The direction vectors are highly correlated with one another because the animal's range of motion is limited. For example, sitting down is characterized by a stereotyped collection of joint angles, and rearing up on the hind legs corresponds to another. Of all possible collections of direction vectors, however, only a small subset are realizable due to physical constraints.

One important degree of freedom remains: poses should be invariant to the direction the animal is facing. Thus, we introduce a latent variable, $\boldsymbol{h} = (h_1, \ldots, h_T)$ where $h_t \in [-\pi, \pi)$, to denote the animal's heading direction in the xy-plane. We model the direction vectors $\boldsymbol{u}$ as conditionally independent given the pose state and the heading direction,

$$u_{t,k} \mid s_t, h_t \sim \mathrm{vMF}\big(\mathrm{R}(h_t)\, \underline{\nu}_{s_t,k}, \kappa_{s_t,k}\big), \tag{8}$$

where $\mathrm{R}(h_t)$ denotes a rotation matrix by angle h_t in the xy-plane, and $\underline{\nu}_{s,k}$ and $\kappa_{s,k}$ denote the mean direction and concentration of a vMF distribution in canonical orientation (i.e. when the heading h_t is zero).

Pose and heading dynamics. Pose states and headings vary over time as well. For simplicity, we treat the headings as independent and uniformly distributed on the circle under the prior. We use a simple Markov model for pose states,

$$s_t \mid s_{t-1} \sim \lambda_{s_{t-1}}, \tag{9}$$

where $\lambda_s \in \Delta_S$ denotes the s-th row of transition matrix $\Lambda \in [0,1]^{S \times S}$.

Robust observation model. Finally, we propose a model of 2D keypoint locations that allows for outliers in the 2D estimates. Let $z_{t,k,c} \in \{0, 1\}$ denote whether observation $y_{t,k,c}$ is an outlier (with 1 denoting an outlier). We allow for different outlier probabilities for each keypoint and camera using the following model,

$$z_{t,k,c} \sim \mathrm{Bern}(\beta_{k,c}) \tag{10}$$

$$\epsilon_{t,k,c} \mid z_{t,k,c} \sim \mathcal{N}(\mu_{k,c,z_{t,k,c}}, \omega^2_{k,c,z_{t,k,c}} I) \tag{11}$$

$$y_{t,k,c} = f_c(x_{t,k}) + \epsilon_{t,k,c}, \tag{12}$$

where $\beta_{k,c}$ denotes the probability that keypoint k will be an outlier on camera c, and $\epsilon_{t,k,c}$ denotes the error between the projected 3D position and the observed 2D keypoint. The outlier variables determine the conditional mean and variance of the error, with outliers having higher variance, $\omega^2_{k,c,1} \gg \omega^2_{k,c,0}$.

The complete model shown in Figure 1d consists of latent variables $(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{s}, \boldsymbol{h}, \boldsymbol{z})$ and data $\boldsymbol{y}$. The model parameters are summarized in Table S1. Next we develop an MCMC algorithm to sample the posterior distribution of latent variables given the 2D position data and parameters, leveraging the conjugacy of this hierarchical model.

4 Algorithm

GIMBAL inherits many of the conjugacy properties of the vMFG model and admits a simple MCMC algorithm for approximate Bayesian inference. Here we develop an MCMC algorithm that combines hybrid Monte Carlo (HMC) (Neal, 2011) and Gibbs updates to target the posterior distribution of latent variables.

Sampling positions First, we sample the 3D positions $\boldsymbol{x}$. These variables do not have a closed form conditional distribution due to the nonlinear projections $f_c(x_{t,k})$, but the unnormalized conditional density $p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{u})$ and its gradients are straightforward to calculate. We use HMC (Neal, 2011) to obtain a transition operator that leaves conditional distribution invariant. Our implementation uses JAX (Bradbury et al., 2018) for automatic differentiation and compilation to CPU, GPU, or TPU. We use 10 leap-frog steps per iteration and adapt the step size during burn-in, following Andrieu and Thoms (2008, eq. 19).

Sampling direction vectors The vMF prior on the mean parameter is conjugate with the Gaussian distribution on positions, just as in the vMFG model of Mukhopadhyay et al. (2019). With direction vector samples proportional to $x_{t,k} - x_{t,\pi(k)}$ for $k = 2, \ldots, K$, and given heading h_t and state s_t, the conditional distribution is vMF with parameters

$$u_{t,k} \mid x_t, s_t, h_t \sim \text{vMF}(\tilde{\nu}_{t,k}, \tilde{\kappa}_{t,k}), \tag{13}$$

where

$$\tilde{\kappa}_{t,k} = \left\| \kappa_{s_t,k} \text{R}(h_t)\, \underline{\nu}_{s_t,k} + \frac{\rho_k}{\sigma_k^2} \left(x_{t,k} - x_{t,\pi(k)} \right) \right\|_2$$
$$\tilde{\nu}_{t,k} = \left(\kappa_{s_t,k} \text{R}(h_t)\, \underline{\nu}_{s_t,k} + \frac{\rho_k}{\sigma_k^2} \left(x_{t,k} - x_{t,\pi(k)} \right) \right) / \tilde{\kappa}_{t,k}.$$

Since the direction vectors are conditionally independent given the pose state and heading, these parameters can be updated in parallel.

Sampling headings Let $\sphericalangle_{\mathsf{xz}}(v)$ and $\sphericalangle_{\mathsf{xy}}(v)$ denote the azimuthal and polar angles, respectively, of a unit vector $v \in \mathbb{S}_2$. Under a uniform prior, the conditional distribution of h_t is,

$$h_t \mid \boldsymbol{u}, s_t \sim \text{vMF}\left(\arctan\left(\frac{\tilde{y}_t}{\tilde{x}_t} \right), \sqrt{\tilde{y}_t^2 + \tilde{x}_t^2} \right), \tag{14}$$

where

$$\tilde{y}_t = \sum_{k=2}^{K} \sin(\sphericalangle_{\mathsf{xz}}(u_{t,k})) \sin(\sphericalangle_{\mathsf{xz}}(\underline{\nu}_{s_t,k})) \sin(\Delta_{t,k})$$
$$\tilde{x}_t = \sum_{k=2}^{K} \sin(\sphericalangle_{\mathsf{xz}}(u_{t,k})) \sin(\sphericalangle_{\mathsf{xz}}(\underline{\nu}_{s_t,k})) \cos(\Delta_{t,k})$$
$$\Delta_{t,k} = \sphericalangle_{\mathsf{xy}}(u_{t,k}) - \sphericalangle_{\mathsf{xy}}(\underline{\nu}_{s_t,k}).$$

Intuitively, $\Delta_{t,k}$ is the angular difference between the given direction $u_{t,k}$ and the canonical prior direction $\underline{\nu}_{s_t,k}$ in the xy-plane. These differences are weighted by the azimuthal angles to determine the conditional distribution of the heading. See Section S3 for a complete derivation.

Sampling pose states Given the direction vectors and headings, the sequence of pose states $\boldsymbol{s}$ is conditionally distributed as,

$$p(\boldsymbol{s} \mid \boldsymbol{u}, \boldsymbol{h}) \propto p(\boldsymbol{s}) \prod_{t=1}^{T} \prod_{k=2}^{K} \text{vMF}\left(u_{t,k} \mid \text{R}(h_t)\, \underline{\nu}_{s_t,k}, \kappa_{s_t,k} \right)$$
$$p(\boldsymbol{s}) = \text{Unif}(s_1) \prod_{t=2}^{T} \Lambda_{s_{t-1}, s_t}. \tag{15}$$

This is a standard hidden Markov model, and we use a standard forward filtering-backward sampling algorithm to sample the conditional distribution.

Sampling outlier indicators Finally, we sample from the conditional distribution,

$$z_{t,k,c} \mid y_{t,k,c}, x_{t,k} \sim \text{Bern}(\tilde{\beta}_{t,k,c}), \tag{16}$$

for

$$\tilde{\beta}_{t,k,c} = \sigma\Big(\sigma^{-1}(\beta_{k,c}) + \log \mathcal{N}(\epsilon_{t,k,c} \mid \mu_{k,1}, \omega_{k,c,1}^2) - \log \mathcal{N}(\epsilon_{t,k,c} \mid \mu_{k,0}, \omega_{k,c,0}^2) \Big),$$

where $\sigma(x) = (1 + e^{-x})^{-1}$ denotes the logistic function and $\sigma^{-1}(\beta) = \log \frac{\beta}{1-\beta}$ denotes its inverse, the logit function.

Setting model parameters GIMBAL's parameters could also be treated in a Bayesian fashion, but we recommend setting them based on a small dataset of "ground truth" 3D positions obtained from triangulation of expert-labeled 2D frames. The direction parameters, $\underline{\nu}_{s,k}$ and $\kappa_{s,k}$, and the transition matrix, Λ, can be found by fitting a hidden Markov model to the ground truth directions, and the observation model parameters $\beta_{k,c}$, $\mu_{k,c,z}$, and $\omega_{k,c,z}^2$ can be found with a mixture of Gaussian model. We describe this process in Section S4.3. The temporal variances η_k^2 can be estimated from ground truth 3D trajectories. We set the paramters ρ_k and variances σ_k^2 by fitting a Gaussian distribution to the distances $\|x_{t,k} - x_{t,\pi(k)}\|$ in the training data.

5 Results

We evaluated GIMBAL's performance on a dataset consisting of six simultaneously recorded video streams of a freely behaving rodent in an open arena (fig. 1a). The subject was affixed with retroreflective markers at 20 keypoints of interest (fig. 1b, c and fig. S2) to collect corresponding ground truth data using 12 infrared motion capture (MOCAP) cameras.

GIMBAL reduces estimation errors We compared GIMBAL to DeepLabCut-3D (DLC-3D) (Nath et al., 2019) and DANNCE (Dunn et al., 2020). DLC-3D uses the median stereo triangulation of 2D keypoints detected by DeepLabCut (DLC), while DANNCE is a volumetric convolutional network that makes 3D predictions from raw video. GIMBAL takes 2D DLC predictions as input, and the model is initialized with the DLC-3D predictions.

We quantified performance with the distance between predicted positions and ground truth MOCAP data, averaged over time. We call this the mean position error (MPE). We calculate MPE both from raw predictions (raw MPE) and from predictions that have been optimally translated and rotated via a rigid Procrustes analysis (RPA-MPE). Note that RPA does not allow scaling, reflection, and other deformations.

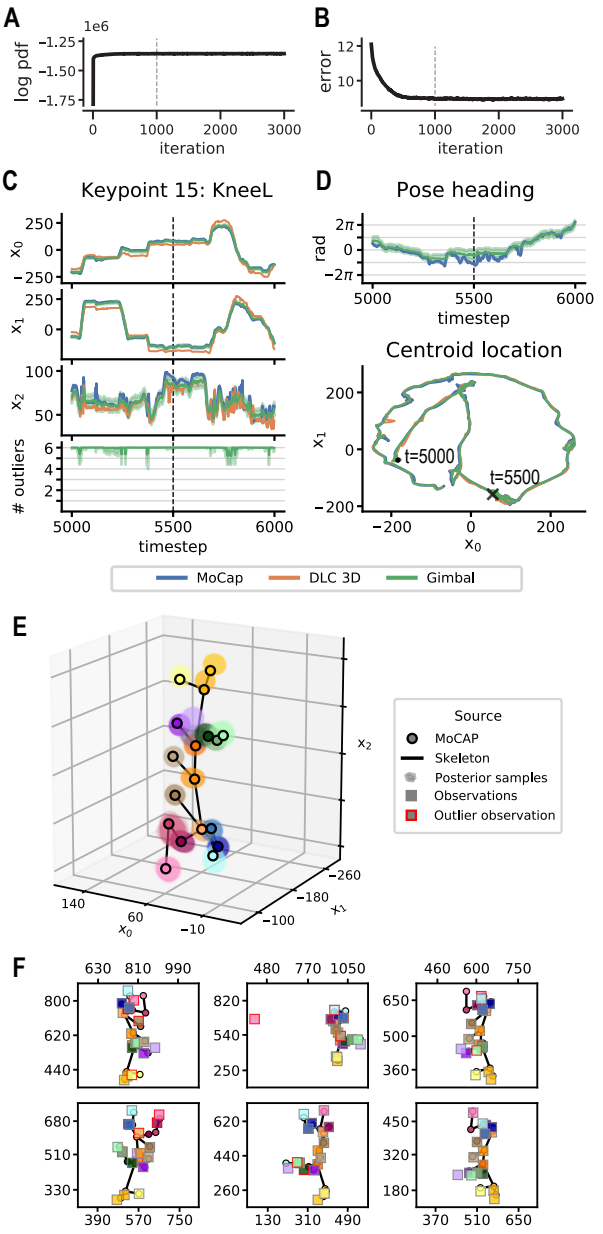

Figure 2: GIMBAL *produces a posterior distribution on 3D pose.* **a.** Log joint probability and **b.** mean position error (MPE) vs. MCMC iterations. Dashed lines indicate burn-in period of $i = 1000$. **c.** Coordinate traces for keypoint 15 from ground truth data (blue), DLC-3D predictions (orange), and GIMBAL predictions ±2 stdv (green). Bottom plot shows the number of inferred outlier observations. **d.** pose heading (true and GIMBAL only) and centroid location in the xy-plane. Dashed lines in **c** and **d** indicate time step $t = 5500$. **e.** 3D pose at time step $t = 5500$ based on MOCAP data (circular markers, black outline) and GIMBAL posterior (point clouds, $n = 2000$). **f.** Corresponding 2D keypoint observations (square markers). Red outlines indicate ≥ 0.70 likelihood of being an outlier. Skeleton segments based on MoCAP positions projected into respective camera plane. Units: px. Marker colors for **e** and **f** correspond to individual keypoints ($K = 20$); see supplementary materials for complete legend. Units for 3D coordinates x_0, x_1, and x_2 axes are in mm.

Table 1: Mean position error (MPE) averaged over all keypoints, for different pose estimation models. Calculated with unmodified predictions (raw) and after applying rigid Procrustes analysis (RPA). Units: mm.

	DLC-3D	DANNCE	GIMBAL
Raw	11.41	9.25	**7.16**
RPA	11.17	7.38	**5.77**

Table 2: Same as Table 1, with results for special submodels of GIMBAL.

	M0	M1	M2	GIMBAL
Raw	16.00	10.71	9.29	**7.16**
RPA	15.40	10.42	8.38	**5.77**

Following convergence (fig. 2a and b), GIMBAL effectively reported individual keypoint positions and outliers (fig. 2c), and subject heading and location (fig. 2d). Figure 2e and f show the predicted 3D pose at a single timestep and corresponding 2D data.

Quantitative results are summarized in Table 1. GIMBAL outperforms DLC-3D by both error measures. On average over all keypoints, the error is close to the 5mm retroreflective marker sized used in collecting the ground truth motion. Since GIMBAL uses the same set of 2D keypoint observations and states with DLC-3D predictions but leverages more structured, prior information, it is expected that GIMBAL achieves higher accuracy. DANNCE can achieve geometric consistency through its 3D convolutional neural network and volumetric representations. Still, GIMBAL achieves an 22.6% improvement in raw MPE and a 21.8% improvement in RPA-MPE over DANNCE.

Figure 3a shows the distribution of the average prediction errors over all keypoints, and Figure 3b shows the errors for four selected keypoints. The methods have similarly high accuracy performance on some keypoints, such as SpineM. Frequently occluded keypoints, however, like ArmR, ElbowL, and KneeL, consistently result in poor DLC-3D performance (fig. 3b). In some cases, such as ArmR, all methods yield poor estimates. GIMBAL is unable to improve the estimation because of more limited training data for this keypoint (see Section S4.3 and Figure S3 in the supplementary materials). In other cases, such as KneeL, GIMBAL infers that, for example, an asymmetric knock-kneed stance is improbable, and leverages its prior distribution on likely directions from the hip to the knee to reduce prediction error by up to 50mm.

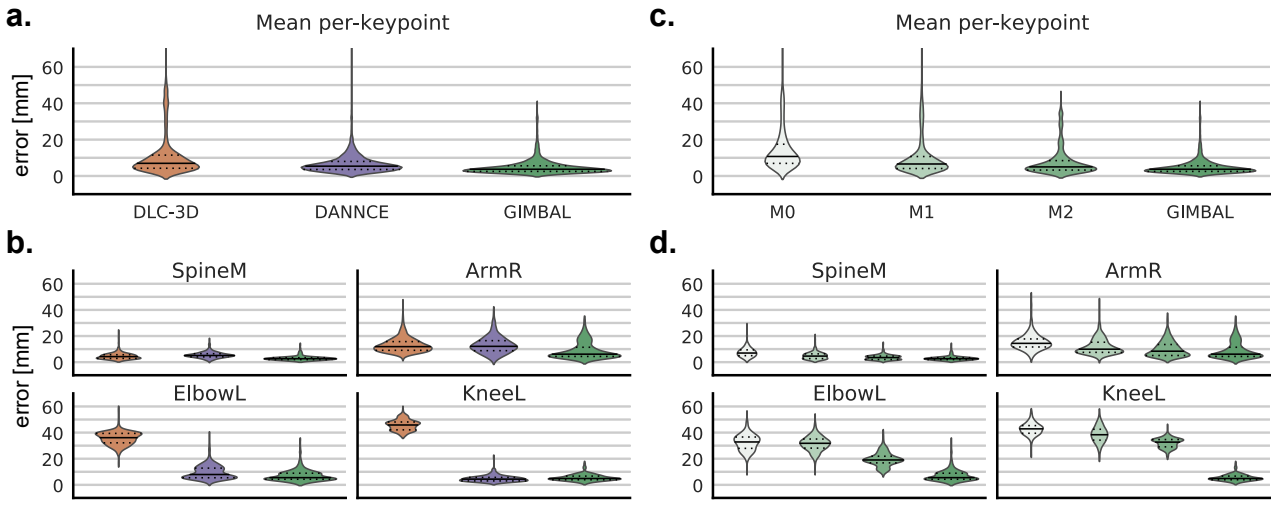

Figure 3: Directional priors make significant contribution to GIMBAL performance. Distribution of **a.** mean position error, averaged across all keypoints, and **b.** per-keypoint mean position error of representative keypoints, for DLC-3D (orange, left), DANNCE (purple, center), and GIMBAL (green, right) predictions. On the right, distribution of **c.** mean position error, averaged across all keypoints, and **d.** per-keypoint mean position error of representative keypoints, for special cases of GIMBAL (i.e. M0, M1, M2). All errors calculated on rigid Procrustes-aligned predictions. Interior lines of violin plot patches indicate respective data quartiles. Median is denote by a solid black line. All means taken over $T = 1000$ timesteps. Units: mm.

Directional priors inform pose estimation In order to elucidate the contribution of directional priors to model performance, we evaluate the performance of special cases of GIMBAL with subsets of latent variables. We refer to these special cases as

- **M0**: Bayesian nonlinear triangulation model. (Includes $\boldsymbol{x}$ only.)
- **M1**: M0 + robust triangulation via a mixture of Gaussians model. (Includes $\boldsymbol{x}$ and $\boldsymbol{z}$.)
- **M2**: M1 + temporal smoothing and distance constraints via a uniform prior on directions. (Includes $\boldsymbol{x}$, $\boldsymbol{z}$, and $\boldsymbol{u}$.)

Parameter settings to implement these special cases are listed in Table S1.

Figure 3c shows the distribution of errors in these nested models. GIMBAL removes a second error mode above 20mm that, upon investigation, was due to certain joints that are commonly mislabeled, like ElbowL and KneeL. While robust observations, distance constraints and temporal smoothing each provide a measure of improvement over their preceding models (e.g. for KneeL, M0: 48.15 ± 4.15 std; M1: 43.12 ± 7.39 std; M2: 36.24 ± 4.40 std), pose priors provide the most significant reduction (GIMBAL: 9.51 ± 3.45 std). Average MPE results are found in Table 2.

Pose state dynamics offer lens into behavioral modeling. GIMBAL's structured prior distributions improved pose estimation, but animal behavior consists of sequences of poses. We conclude by studying how GIMBAL's pose state dynamics offer views into animal behavior.

Figure 4a shows the posterior probability of each state over time frame. The sequences of pose states shown in the insets reflect rearing up on the hind legs and perambulation (i.e. walking). No states in the range $[0, 20]$ were exhibited in this 33 second window. These pose dynamics and their connection to behavior emerged after permuting states based on physical positioning. Specifically, pose states were grouped into coarse categories and then sorting based on uprightness. Figure 4b provides an example mean pose from each category.

The transition matrix (fig. 4c) captures some of these pose state dynamics. For example, rearing up consists of a serial transition from states in the range $[21, 62]$ through $[63, 84]$ to $[85, 119]$, and reversed for rearing down. This can be seen in the blocks composing the lower quadrant of Figure 4c and d. Perambulation is a periodic sequence of pose states in the range $[21, 62]$. Since this transition matrix encodes the probabilities for all possible pose state dynamics, the matrix structure associated with a weakly cyclical Markov chain is diffuse, but this structure is roughly observable in Figure 4d.

6 Discussion

We presented GIMBAL, a hierarchical von Mises-Fisher-Gaussian model for animal pose estimation. Our model improves 3D pose estimation results from multiview keypoint observations over two existing animal pose estimation methods. GIMBAL additionally provides statistically- and geometrically-relevant uncertainty estimates that are lacking in current deep learning-based approaches.

We systematically examined the contributions of each model component, and showed that the hierarchical von Mises-Fisher-Gaussian model provided a significant source of improvement in keypoint prediction accuracy. This model reduction (M0, M1, and M2) also provides a convenient framework in which to compare GIMBAL to other recent works. For example, state-of-art methods incorporating temporal smoothing and distance constraints (e.g. Günel et al. (2019), Karashchuk et al. (2020)) are analogous to M2. GIMBAL concisely captures the manifold constraints on skeletal keypoint positions, leading to better performance with a simple MCMC algorithm.

GIMBAL's can be readily adapted to observations other than 2D keypoint detections—for example, image features or network heatmaps with multiple candidates. The hierarchical von Mises-Fisher-Gaussian model is agnostic to the choice of likelihood, as long as it is amenable to automatic differentiation for HMC.

Finally, we examined how behavior is captured by the sequence of inferred pose states. Previous methods relied on states derived from imprecise sources such as image features (Wiltschko et al., 2015). Our representation consists of a collection of joint angles, which can accurately recover individual keypoint positions. Additionally, this representation provides a basis for sharing pretrained pose states and dynamics across animals with similar biomechanics and behaviors.

Future work could also explore continuous representations of pose states and dynamics — in the spirit of hierarchical GP-LVMs (Kearney et al., 2020), but with spherical manifold constraints—to enable better action recognition and behavioral modeling. Likewise, we expect modular representations of pose, which capture the configurations of subsets of keypoints and the relationships between those subsets, could lead to new understanding of postural dynamics.

GIMBAL leverages recent advances in spherical manifold learning with hierarchical von Mises-Fisher-Gaussian models to improve upon existing techniques in animal pose estimation. Its latent state representations of pose and postural dynamics offer promising new building blocks for animal behavioral modeling.

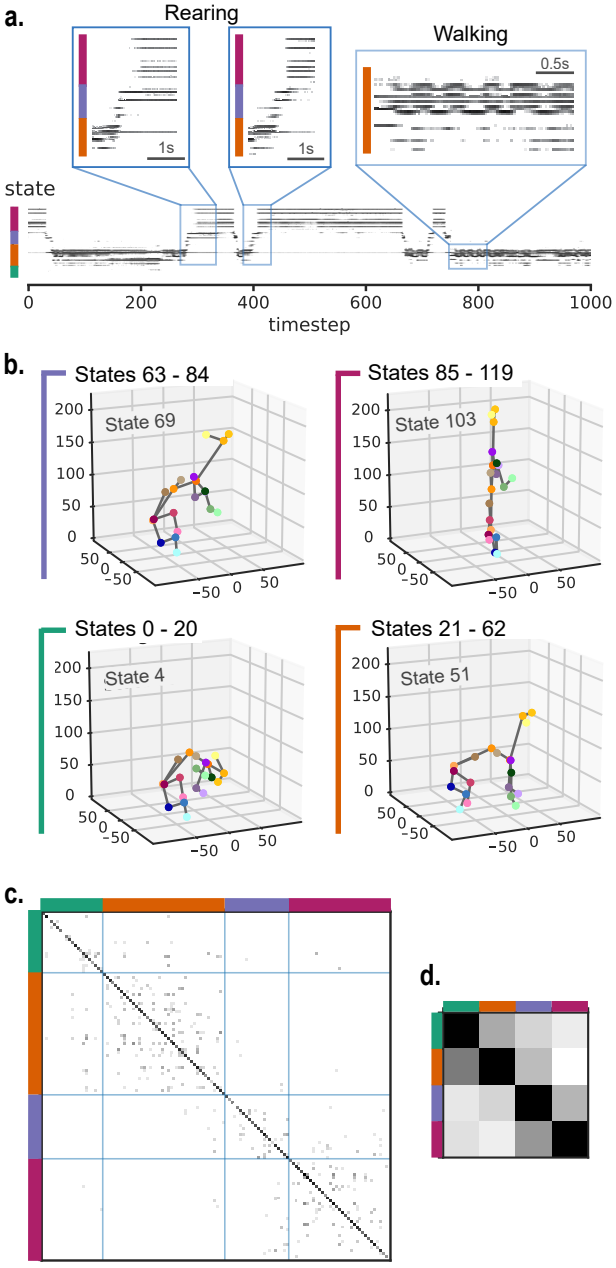

Figure 4: Pose state dynamics are descriptive of observed behaviors. **a.** Inferred state trajectory over time, for $T = 1000$ time bins and $S = 120$ discrete pose states. Magnified sections draw attention to stereotyped behaviors evident in the state trajectory, such as rearing up and walking. No states in range $[0, 20]$ (green) were exhibited during this data sequence. Latent pose states were permuted into four coarse categories, and then sorted within category by z-coordinate of HeadB keypoint. **b.** Example pose for each of the state ranges. Each pose is drawn using mean direction under the prior. **c.** Empirical transition matrix $\Lambda \in [0, 1]^{S \times S}$. Light blue lines delineate category boundaries. **d.** Summary of category-level transitions based on **c**. All frequencies plotted on a logarithmic scale and normalized to the minimum and maximum values of the matrix.

References

S. Amin, M. Andriluka, M. Rohrbach, and B. Schiele. Multi-view pictorial structures for 3d human pose estimation. In *The British Machine Vision Conference (BMVC)*, volume 1, 2013.

C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Statistics and Computing*, 18(4):343–373, 2008.

P. C. Bala, B. R. Eisenreich, S. B. M. Yoo, B. Y. Hayden, H. S. Park, and J. Zimmermann. Automated markerless pose estimation in freely moving macaques with openmonkeystudio. *Nature Communications*, 11 (1):1–12, 2020.

V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3d pictorial structures for multiple human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1669–1676, 2014.

G. J. Berman, D. M. Choi, W. Bialek, and J. W. Shaevitz. Mapping the stereotyped behaviour of freely moving fruit flies. *J. R. Soc. Interface*, 11(99), Oct. 2014.

B. Biggs, T. Roddick, A. Fitzgibbon, and R. Cipolla. Creatures great and SMAL: Recovering the shape and motion of animals from video. Nov. 2018.

J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, and S. Wanderman-Milne. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

R. E. Brown and S. Bolivar. The importance of behavioural bioassays in neuroscience. *J. Neurosci. Methods*, 300:68–76, Apr. 2018.

M. Burenius, J. Sullivan, and S. Carlsson. 3d pictorial structures for multiple view articulated pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3618–3625, 2013.

A. I. Dell, J. A. Bender, K. Branson, I. D. Couzin, G. G. de Polavieja, L. P. J. J. Noldus, A. Pérez-Escudero, P. Perona, A. D. Straw, M. Wikelski, and U. Brose. Automated image-based tracking and its application in ecology. *Trends Ecol. Evol.*, 29(7):417–428, July 2014.

T. Dunn, J. Marshall, K. Severson, D. Aldarondo, D. Hildebrand, S. Chettih, W. Wang, A. Gellis, D. Carlson, D. Aronov, W. Freiwald, F. Wang, and B. Ölveczky. DANNCE: 3-dimensional aligned neural network for computational ethology. *Nature Methods*, (In press), 2020.

P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *Int. J. Comput. Vis.*, 61(1):55–79, 2005.

A. Gomez-Marin, N. Partoune, G. J. Stephens, and M. Louis. Automated tracking of animal posture and movement during exploration and sensory orientation behaviors. *PLoS One*, 7(8):e41642, Aug. 2012.

J. M. Graving, D. Chae, H. Naik, L. Li, B. Koger, B. R. Costelloe, and I. D. Couzin. DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning. Sept. 2019.

S. Günel, H. Rhodin, D. Morales, J. Campagnolo, P. Ramdya, and P. Fua. DeepFly3D, a deep learning-based approach for 3D limb and appendage tracking in tethered, adult drosophila. *Elife*, 8, Oct. 2019.

R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

K. Hornik and B. Grün. On maximum likelihood estimation of the concentration parameter of von mises–fisher distributions. *Computational Statistics*, 29:945–957, 2014.

H. Jhuang, E. Garrote, J. Mutch, X. Yu, V. Khilnani, T. Poggio, A. D. Steele, and T. Serre. Automated home-cage behavioural phenotyping of mice. *Nat. Commun.*, 1:68, Sept. 2010.

P. Karashchuk, K. L. Rupp, E. S. Dickinson, E. Sanders, E. Azim, B. W. Brunton, and J. C. Tuthill. Anipose: a toolkit for robust markerless 3D pose estimation. May 2020.

S. Kearney, W. Li, M. Parsons, K. I. Kim, and D. Cosker. Rgbd-dog: Predicting canine pose from rgbd sensors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8336–8345, 2020.

J. W. Krakauer, A. A. Ghazanfar, A. Gomez-Marin, M. A. MacIver, and D. Poeppel. Neuroscience needs behavior: Correcting a reductionist bias. *Neuron*, 93 (3):480–490, Feb. 2017.

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

M. Mukhopadhyay, D. Li, and D. B. Dunson. Estimating densities with nonlinear support using Fisher-Gaussian kernels. *arXiv preprint arXiv:1907.05918*, 2019.

S. Musall, A. E. Urai, D. Sussillo, and A. K. Church-

land. Harnessing behavioral diversity to understand neural computations for cognition. *Curr. Opin. Neurobiol.*, 58:229–238, Oct. 2019.

T. Nath, A. Mathis, A. C. Chen, A. Patel, M. Bethge, and M. W. Mathis. Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nat. Protoc.*, 14(7):2152–2176, July 2019.

R. Neal. MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, editors, *Handbook of Markov Chain Monte Carlo*, pages 113–162. CRC Press, May 2011.

G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Harvesting multiple views for marker-less 3D human pose annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6988–6997, 2017.

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

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

A. Wu, E. Kelly Buchanan, M. Whiteway, M. Schartner, G. Meijer, J.-P. Noel, E. Rodriguez, C. Everett, A. Norovich, E. Schaffer, N. Mishra, C. Daniel Salzman, D. Angelaki, A. Bendesky, The International Brain Laboratory, J. Cunningham, and L. Paninski. Deep graph pose: a semi-supervised deep graphical model for improved animal pose tracking. Aug. 2020.

Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1385–1392, 2011.