We have read and appreciate all comments, due to page limit we will only address a subset of questions/concerns.
R1: The approximation of the normalizing factor is inelegant Since submitting the paper we have implemented a normalizing factor based on numerical integration of the equations (14) and (15) in [11]. The planned method uses 512 trapezoid integrals with float32 precision. We plan to use this new method for the camera ready, remove all mentions of the old method and rerun all experiments. We do not expect this will significantly change our performance.

R1: Can you evaluate the coarseness of the approximation? We use the integration method with 2^{14} trapezoids and float128 precision as the true function.

By sampling 1000 random singular values with an L2 norm < 50 we get the following results. The relative error of the forward pass $|log(\hat{a}(F)) - log(a(F))|/log(a(F))$ has an average error of 0.07 with the old method. This error for the integration method (512 trapezoids) is $\approx 10^{-5}$. The norm of error for the backward pass $||\nabla_F log(\hat{a}(F)) - \nabla_F log(a(F))||_F$ is for the old method on average $\approx 10^{-1}$. With the integration method (512 trapezoids) it is $\approx 10^{-3}$.

R4: Approximation $\implies$ loss not necessarily convex: This is true. Hopefully though as we are approximating a convex loss this still greatly helps with the convergence and stability of training. Also any standard approximation (look-up table, rbf as in [5], horner's method etc.) could potentially affect the convexity.

R1: approximation $\implies$ estimated R is not in $SO(3)$? The approximation affects the predicted F and thus the R we estimate but not the property of it being in $SO(3)$. The estimated R, obtained from F via the steps described in section 3.1, is guaranteed to be in $SO(3)$.

R1: Do the model get stuck in bad local optima for classes with symmetry or visual similarity for different poses, especially early in the training As described in line 219 - 229 What we do observe is that the network early on gives large variance for the axis which it cannot identify. This is not a local minima for the network outputs, since the optima is to predict the correct pose confidently and the loss is (approximately) convex. We observe a reasonably stable reduction in training loss (fig 2 in supplementary) this indicates that we do not get stuck in local minimas for long, if at all.

R1: What do you mean by solving the problem of non-positive semidefinite matrices? ${}_1F_1$ is defined for positive semidefinite matrices, see [2]. It can be extended to non positive semidefinite matrices as in appendix eq. (26) or in [2]. **Clarify "proper rotation matrix"?** A proper rotation matrix is ON and $det(R) = 1$

R2: Is the posterior F well aligned with the empirical error? We have not investigated this rigorously, but anecdotally this seems to be the case for large errors. For small errors it is probably not the case since we over regularize (line 115-117 in paper), this disincentives the network from returning low variance outputs.

R2: Would using only the $tr(F^T R)$ as a loss work? No. Without regularization it is possible to get arbitrarily low loss with high average angle errors by outputting large magnitude F's. Using other regularizers such as $||F||_F^2$ could work, but we have used a regularizer with a probabilistic motivation.

R4: Could you do an ablation with more/all other competing losses? Ideally yes, but we do think that the current ablation shows that even when removing data augmentations and our warp preprocessing we outperform other methods. We think this is strong evidence that our method is significantly better. Doing ablations can be problematic due to availability of code and sensitivity to tuning.

R2: Do you use pytorch SVD for backpropagation? Yes we mainly use pytorch svd. We do some custom handling to avoid instability from using $det(U^T V)$

R2: Could you compare to [5]? Empirically we evaluated on UPNA for this reason. (line 197-199) Their code was not available at the time of writing. Their loss is discontinous and non convex due to Gram-Schmidt.

R2: Would any probabilistic approach perform better [than Current SOTA] Not necessarily, [21] (probabilisitc) is outperformed by [16] (not probabilistic) which is outperformed by ours (table 1).

R4: Instead of using the mode could you use an Bayes estimator for some distance? Due to symmetry the two coincide for many reasonable distances. see Theorem 2.3 in [11] for one example.

R1, R4: The method has a problem with rotation symmetric objects. We developed this method with identifiable orientations in mind. We noticed some classes had rotation symmetries and we thought our method had interesting behaviour for these objects. Another distribution should be used when estimating orientation of symmetric objects.

R4: The loss have the properties (Convexity, rotation invariance, bounded gradient etc.), which properties are present in competing methods? We did not investigate this thoroughly. In general these properties are the exception not the rule. We believe methods which use quaternions(discontinuity), Euler angles(rotation invariance), classification(rotation invariance) or Gram-Schmidt (discontinuity) will lose at least one of these properties.