We would like to thank you and appreciate your clear review, important points and invested time. Before going through your comments, we want to inform you that we have added the results of another high dimensional experiment, visual odometry task. Please note that these results were almost ready at the time of the submission, but due to the page limit we decided to cover two high dimensional experiments, single-double pendulum, and two low dimensional experiments, Lorenz and NCLT. But, because of the concerns of some reviewers about higher dimensional experiment results, we have merged single-double pendulum experiment and added visual odometry results in the rebuttal version.

We try to clarify point by point based on your comments and feedback.

Significance. Effectively, this method feels much like the Recurrent Kalman Network, where some approximations are relaxed by using convolutions for the covariance matrix transitions and the direct mapping to the KG and SG. It is not surprising that, given that the ground truth targets are available, this results in a slightly better performance

R1.

1- The utilized factorization in the Recurrent Kalman Network results in a biased estimate of Covariance matrix. We already knew that the covariance matrix possibly is sparse, but there is no guarantee that these elements are located at the side diagonal parts of the covariance matrix as they assumed in the RKN.

2- In the GIN, There is no restriction for latent state dimension, n, and the observation dimension, m, i.e. dimension of $\mathbf{x}$ and $\mathbf{w}$ in the paper. It is possible to set $m = n$, for the cases where the observation is considered as a noisy measured states, e.g. Lorenz Attractor experiment. While one can divide the latent state space into two parts $n = 2m$, e.g. NCLT experiment. By considering $n = 3m$, one can include the acceleration information in the states, while only positional noisy measurements are obtained, e.g. what we have done in our high dimensional experiments: single-double pendulum and visual odometry. Totally, the size of the latent state is independent of the observation, while in the RKN it is restricted to the $n = 2m$. By doing so, conducting many experiments like Lorenz Attractor is not feasible with the RKN.

3- In the RKN, their assumption is based on the lack of the dynamics and their proposed model can not utilize known or at least partially known dynamics. Our GIN handles both, where the dynamics are known, e.g. Lorenz Attractor and NCLT, and unknown dynamics, e.g. single-double pendulum and visual odometry. One of the reasons that we also include the experiments with the low dimensional observations and known dynamics, e.g. Lorenz attractor and NCLT, is to show that the GIN is able to handle this case. Although, it is possible to conduct the experiment without dynamics knowledge, but it has negative effects on the performance (we also perform this type of the setting for both Lorenz and NCLT experiments, you can find the code in "unknown dynamics folder").

The paper feels convoluted. I have the feeling that sometimes function arguments are missing. For example, F and H are functions of both Q_t and $\mathbf{x}^+_{t-1}$ right?

R2.

From section 3.1, F_t and H_t both are function of $\mathbf{x}_{t-1}^+$ with this equation: $\mathbf{F}_t = \sum_{k=1}^K \alpha_t^k \mathbf{F}_t^k, \quad \mathbf{H}_t = \sum_{k=1}^K \alpha_t^k \mathbf{H}_t^k$. Here, $\alpha^k(\mathbf{x}_{t-1}^+)$ is the output of a DNN with softmax activation and ($F_t^{1:K}$, $H_t^{1:K}$) are K different trainable matrices.

About the function arguments, due to some lengthy arguments, including every argument make the formulas messy. So once functions are defined, the explicit arguments are omitted for clarity.

Besides from $\mathbf{F}$, learning the process noise Q_t is conducted separately. Based on section 3.2, we propose three different ways for learning Q_t:

1- From eq. (23), we write down Q_t as a function of F_t, where F_t is a function of $\mathbf{x}_{t-1}^+$. By this notation, we assumed the learned F_t can also includes the effects of Q_t: re-write eq. (3) as eq. (4).

2- Q_t can be directly derived as a function of $\mathbf{x}_{t-1}^+$. For this method, we use another small DNN with the input of $\mathbf{x}_{t-1}^+$ and output of a lower triangular matrix L_t with positive diagonal elements, and calculate $Q_t = L_t L_t^T$. By this solution, we use original eq. (3).

3- Q_t can be written as a recursive function of F_t and Q_{t-1}, as we derived in eq. (21). In this method, we use a small GRU network with the input of F_t and output of a lower triangular matrix L_t with positive diagonal elements, and calculate $Q_t = L_t L_t^T$. By this solution, we use original eq. (3). The results of these three methods are in section A.9 , where modeling Q_t with GRU cell slightly improves the overall MSE of single-double pendulum experiments.

Why is the likelihood introduced as $p_\theta(\mathbf{y}|\mathbf{x}_\theta, \mathbf{\Sigma}_\theta)$? What is $\mathbf{y}$ here, and θ?

R3.

For the consistency of the notations in the paper, we have modified this notation to $\mathcal{L}_s := \log \prod_{t=1}^T p(\mathbf{s}_t|\mathbf{o}_{1:T})$ for the state estimation task, i.e. $\mathbf{s}_t$ is equal to $\mathbf{o}_t^+$ in the figure 2. Please see lines 205-208 in the rebuttal version.

Figure 2 is too complex. Maybe it can be split up into sub-components.

R4. We have proposed 3 modifications for the figure 2.

1- An overall structure 1, without subfigures in this link

2- An overall structure 2, without subfigures in this link

3- An overall structure with three separated subfigures in this link

Currently, we are using the first option. But if you find other options better, we can switch.

Originality. The parameterization of KG and SG being the main selling point of the paper does not feel like an original take on the problem.

R5.

We agree that the originality of the paper may be considered as several existing ideas, e.g. 1-using Recurrent cells for estimating Kalman/Smoothing gain 2- exploiting the denoising abilities of $f(.)$ and $h(.)$ by implementing them with DNNs 3- modeling the unknown dynamics with DNN 4- using the dynamics when they are available 5- Handling low dimensional data 5-handling high dimensional data, etc.

But, to the best of our knowledge, there is not a model which is able to handle all of the mentioned situations. For example in the KVAE and RKN, they are just able to model high dimensional data without known dynamics with latent size restriction. Or in KalmanNet, HybridGNN and the paper of Ruhe et al., their model is just able to handle low dimensional data with the knowledge of dynamics

The paper proposes a series of parameterizations and then suddenly specifies a (pseudo-)likelihood that is optimized. It seems to me that the paper parameterizes $p(\mathbf{s}_t|\mathbf{o}_{1:T})$ in a complicated manner. However, since this distribution is partly directly parameterized anyway (through the learned Kalman gain matrices), can't we directly parameterize all of it using a flexible model entirely and rid us of the filtering and smoothing steps?

R6.

Firstly, let us compare the GIN with a variational-inference based model in the case of parameterization.

1- In a variational-based approach, usually $p(\mathbf{s})$ is parameterized as: $\log \int p(\mathbf{s}, \mathbf{w}, \mathbf{x}) d\mathbf{w} d\mathbf{x} \geq \mathbb{E}_{q(\mathbf{x},\mathbf{w}|\mathbf{s})} \left[\log \frac{p(\mathbf{s}|\mathbf{w}) p(\mathbf{w}|\mathbf{x}) p(\mathbf{x})}{q(\mathbf{w},\mathbf{x}|\mathbf{s})}\right] = \mathbb{E}_{q(\mathbf{w}|\mathbf{s})} \left[\log \frac{p(\mathbf{s}|\mathbf{w})}{q(\mathbf{w}|\mathbf{s})} + \mathbb{E}_{p(\mathbf{x}|\mathbf{w})} \left[\log \frac{p(\mathbf{w}|\mathbf{x}) p(\mathbf{x})}{p(\mathbf{x}|\mathbf{w})}\right]\right] = \mathcal{L}_s$ where they use $q(.)$ as an approximation of the posterior and then try to estimate lower bound with MC sampling and maximize it.

2- Despite variational method, in the GIN we try to directly maximize a pseudo-likelihood of $\mathbf{s}$, which enables us to perform end-to-end optimization. So to answer your question : *What is the advantage of Kalman filter parameterization over directly parameterizing the posterior?*, the reasons that we introduce $p_\gamma(\mathbf{w}, \mathbf{x}) = \prod_{t=1}^{T} p_{\gamma_t(\mathbf{x}_{t-1}^{+})}(\mathbf{w}_t|\mathbf{x}_t).p(\mathbf{x}_1) \prod_{t=2}^{T} p_{\gamma_t(\mathbf{x}_{t-1}^{+})}(\mathbf{x}_t|\mathbf{x}_{t-1})$ notation and Kalman Filter parameterization, are to

(i)- present how the data flows in the filtering/smoothng steps and how finally $(\mathbf{x}_t^{+}, \mathbf{\Sigma}_t^{+})$ are obtained, where $p(\mathbf{w}_t|\mathbf{o}_t) = \mathcal{N}(\mathbf{w}_t; \text{enc}_{mean}(\mathbf{o}_t), \text{enc}_{sigma}(\mathbf{o}_t))$ and $p_{\gamma_t(\mathbf{x}_{t-1}^{+})}(\mathbf{w}_t|\mathbf{x}_t) = \mathcal{N}(\mathbf{w}_t|\mathbf{H}_t\mathbf{x}_\mathbf{t}, \mathbf{R}_t)$
and $p_{\gamma_t(\mathbf{x}_{t-1}^{+})}(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_\mathbf{t}; \mathbf{F}_t\mathbf{x}_{t-1}, \mathbf{Q}_t)$ and finally $p_{\gamma_t(\mathbf{x}_{t-1}^{+})}(\mathbf{x}_t|\mathbf{w}_t) = \mathcal{N}(\mathbf{x}_t; \mathbf{x}_t^{-} + \mathbf{K}_t[\mathbf{w}_t - \mathbf{H}_t\mathbf{x}_t^{-}], \mathbf{\Sigma}_t^{-} - \mathbf{K}_t[\mathbf{H}_t\mathbf{\Sigma}_t^{-}\mathbf{H}_t^T + \mathbf{R}_t]\mathbf{K}_t^T) = \mathcal{N}(\mathbf{x}_t; \mathbf{x}_t^{+}, \mathbf{\Sigma}_t^{+})$

(ii)- We can present the procedure of obtaining Kalman Gian(similarly smoothing gain) based on this parameterization, and propose a meaningful structure for modeling Kalman Gain (smoothing gain), i.e. GRU^{KG} and GRU^{SG} in the paper.

(iii)- Showing how the dynamic parameters, i.e. $(\mathbf{F}, \mathbf{H})$, can be written as a function of last posterior $\mathbf{x}_{t-1}^{+}$ so that we can justify the idea of constructing dynamic parameters based on the last posterior with the Dynamics Network.

(iv)- How the data finally construct the pseudo-likelihood $p(\mathbf{s}_t) = \mathcal{N}(\mathbf{s}_t|\mathrm{dec}_{mean}(\mathbf{x}_t^+), \mathrm{dec}_{sigma}(\mathbf{\Sigma}_t^+))$.

So we think it is feasible to directly parameterize $p(\mathbf{s}_t|\mathbf{o}_{1:T})$ to your question: *can't we directly parameterize all of it using a flexible model entirely?*. Because we are trying to maximize the pseudo-likelihood, but we found it useful to include Kalman Parameterization to let the author have some intuition about the proposed structure and data-flow. With this explanation, if you still find the notations confusing, please let us know to modify.

Further, with many nonlinear parameterizations, is it still reasonable to use Kalman filter updates, as Gaussianity is probably lost?

R7.

There is one intuition, based on which, we still can assume the Gaussianity remains. Lets say we have a dynamic system of the form: $\mathbf{x}_t = f(\mathbf{x}_{t-1}) + q_t$ and $\mathbf{w}_t = h(\mathbf{x}_t) + r_t$. Then by applying linearization, the system can be written as $\mathbf{x}_t = \mathbf{F}_t(\mathbf{x}_{t-1})\mathbf{x}_{t-1} + q_t$ and $\mathbf{w}_t = \mathbf{H}_t(\mathbf{x}_{t-1})\mathbf{x}_t + r_t$. If we conduct filtering parameterization under Gaussianity assumption as $p(\mathbf{x}_t|\mathbf{w}_{1:t})$, then the posteriors are $(\mathbf{x}_t^- + \mathbf{K}(\mathbf{H_t})_t[\mathbf{w}_t - \mathbf{H}_t\mathbf{x}_t^-], \mathbf{\Sigma}_t^- - \mathbf{K}(\mathbf{H_t})_t[\mathbf{H}_t\mathbf{\Sigma}_t^-\mathbf{H}_t^T + \mathbf{R}_t]\mathbf{K}(\mathbf{H_t})_t^T)$. We use $\mathbf{K}(\mathbf{H_t})_t$ notation to show that the Kalman Gain is function of $\mathbf{H}_t$ and accordingly $\mathbf{x}_{t-1}$. In order to keep Gaussianity, we have to provide a close approximation of the posterior to guarantee that the posteriors are still following Gaussian, meaning that if trained $(\mathbf{F}_t, \mathbf{H}_t)$ (with the Dynamics network in the paper) and trained Kalman Gain (GRU^{KG} in the paper) can provide a good approximation of the mentioned posteriors such that $GRU_t^{KG} \approx \mathbf{K}(\mathbf{H_t})_t$, we still can assume Gaussianity. During the training phase, by imposing Gaussian distributin over $p(\mathbf{s}_t)$ we can force $(\mathbf{x}_t^+, \mathbf{\Sigma}_t^+)$ to still stay Gaussian. In other words, after convincing number of epochs, we expect $(\mathbf{x}_t^+, \mathbf{\Sigma}_t^+)$ to be almost Gaussian.

To show this, we conduct some visualization to ensure the Gaussianity assumption of the posteriors is realistic and using Kalman parameterization is reasonable (note that the same reasoning for the smoothing parameterization is feasible).

1- In the pendulum experiment, at $t = 100$, we generated samples of x position of the joint of the single pendulum from the smoothened distribution, $f(x1_{100}|\mathbf{w}_{1:150})$, y position from smoothened distribution, $f(x2_{100}|\mathbf{w}_{1:150})$, and (x,y) from the joint distribution. Then we fit a density on the generated samples to indicate the Gaussianity. Additionally, this visualization shows the effectiveness of the GIN in reducing the uncertainty of the estimates compare to LGSSM and KVAE. Please see figures or the paper long version page 18.

2- The same visualization for the double pendulum. Figures or paper long version pages 19-20.

3- The same visualization for the visual odometry experiment. Figures or paper long version page 21.

Similarly, for a Bernoulli likelihood. This is also not clear from the paper.

R8.

For the image imputation task, in addition to the pseudo-likelihood for inferring the states, we add the reconstruction pseudo-likelihood for inferring images by using Bernoulli distributions as

$\mathcal{L}_{\mathbf{i}} := \log \prod_{t=1}^{T} p(\mathbf{i}_t|\mathbf{o}_{1:T})$, i.e. the decoder maps both state $\mathbf{s}_t$ and image $\mathbf{i}_t$: $\mathbf{o}_t^+ = [\mathbf{i}_t, \mathbf{s}_t]$. Then we consider log-likelihood as

$$\mathcal{L}(\mathbf{o}_{1:T}^+) = \mathcal{L}(\mathbf{s}_{1:T}) + \lambda \sum_{t=1}^{T} \sum_{k=0}^{D_o} \mathbf{i}_t^{(k)} \log\big(\mathrm{dec}_k(\mathbf{x}_{\mathbf{t}|\mathbf{T}})\big) + \big(1 - \mathbf{i}_t^{(k)}\big) \log(1 - \mathrm{dec}_k(\mathbf{x}_{\mathbf{t}|\mathbf{T}})).$$

$\mathrm{dec}_k(\mathbf{x}_t)$ defines the corresponding part of the decoder that maps the k-th pixel of $\mathbf{i}_t$ image and λ constant determines the importance of the reconstruction.

Related to the previous bullet point: this paper in the recent ICLR proceedings seems highly relevant and has overlapping locally linear approximations but is not cited. In this work, it is clear what the assumptions, parameterizations, and approximations are. I would suggest the authors similarly clarify these things.

R9.

Thanks for the related paper and your recommendation, we have cited this paper and also included some of the parameterization similarly in lines 131-135 and appendix A.1.

Why are $\mathbf{F}_t$ and $\mathbf{H}_t$ parameterized as softm ax-weighted combinations instead of directly by a neural network?.

R10.

We globally learn K basic state transition and emission matrices, $\mathbf{F}^k$ and $\mathbf{H}^k$ and interpolate them based on the information from the latent state $\mathbf{x}$. This formulation can be interpreted as a soft mixture of K different dynamic parameters whose time-invariant matrices are combined using the time-varying weights as the output of a DNN. Intuitively, each of k sets of $\mathbf{F}^k, \mathbf{H}^k$ models different dynamics, that will dominate when the corresponding element of the dynamics network is high. Additionally, this formulation can help us to introduce a loss term which tries to increase the distance of each pair of $\mathbf{F}^k, \mathbf{H}^k$ set. We found this method as a potential solution to prevent the model goes through the mode collapse.

Why is Satorras et al. left out of experiment 4.2.2. even though they also have this experiment, and the authors do include them in 4.2.1. which is also taken from that paper?

R11.

The result of this experiment is also included in the new version of the paper. Please refer to NCLT experiment section.

The authors briefly mention as a limitation that the model has not been tested on complex real-world videos. I believe there are many more (implicit) approximations (e.g., Gaussianity) made throughout the paper that limit the model. This is not clear from the text.

R12.

The new included visual odometry experiment can be considered as a higher dimensional experiment with further complexities. We hope that the reasons and visualizations mentioned in **R7** address your concerns and will be happy to answer any further questions.