We thank all reviewers for their feedback! We are encouraged to see that the reviewers found our paper "clear, interesting, and novel" (R1, R2), "investigating an important issue" (R1, R3, R4), well-placed "into the related literature up to some very recent works" (R1), and "essential reading for anyone using RNVP / Glow-style architectures" (R2).

In the paper we study inductive biases of coupling layer based normalizing flows and show how they influence OOD detection — which types of data are assigned low and high likelihood. We identify mechanisms through which flows learn to improve likelihoods simultaneously on all structured images, even with semantics unrelated to training data. We believe that our analysis of flow's biases will be valuable for both DGM and OOD detection research communities.

R1. Thank you for your positive review! You are correct that our analysis is specific to coupling layer based normalizing flows. However, we believe that the intuition presented in the Sec. 4 of the paper can be translated to other models. We hope that our work will inspire similar investigations into the inductive biases of other DGMs.

R2. Thank you for your positive and detailed feedback! We agree that likelihood is not necessarily the right measure to detect OOD data, and we are not advocating for it. In this paper we study the question of how flows assign likelihood to data including OOD inputs. While typicality explains how in some particular cases OOD detection with likelihood fails (e.g. for high dimensional Gaussian), we argue that it does not by itself explain the likelihood assignment of more complex models like normalizing flows. We will clarify this point and add a discussion in the updated version.

R3. Thank you for your detailed feedback. While we agree with some of the points you raise, we respectfully disagree with your assessment. We believe that our empirical results would be valuable to the community and our narrative is consistent with the observations. We will clarify the limitations of our observations in the updated version of the text.

Figure 1: dz_i/dx

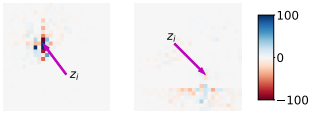
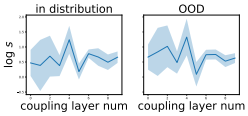


Figure 2: $\log s$ by layer



1) We agree that flows are likely learning *some* information about the semantics, but it has little effect on the likelihood. Our paper identifies mechanisms through which flows can improve likelihood of the data regardless of the semantics. We will clarify that we are not arguing that flows do not learn semantics. Flows produce samples that resemble the training data semantically; however, how exactly they generate samples has to the best of our knowledge not yet been studied deeply, e.g., it is not clear to what extent flows memorize training data, how well they generalize and consequently whether they learn semantics.
2) The main purpose of Section 5 is to demonstrate that the latent representations learned by the flows can be interpreted and we can observe the edges of the original images in the latent space. We agree that on its own the presented results do not prove that flows do not learn the semantic information about the data. However, this observation is novel and at least partly contradicts the intuition that flows perform very complex high-dimensional transformation of the data; it provides motivation for the subsequent sections. Averaging dequantization allows us to denoise the latent representations to more clearly demonstrate that the latent representations contain the edges of the original inputs. We will clarify these points further in an updated version of the paper. • We performed the experiment you suggested and visualized the contributions of the different pixels to the Jacobian. For most latent coordinates the receptive field is limited to the neighbouring pixels (as in Fig. 1 left) but some coordinates are affected by longer range dependencies (Fig. 1 right). Thank you for the idea for this experiment, we will include a detailed description in the updated version of the paper.
3) Note that in Fig. 3 of the paper we are using a flow with two coupling layers, not the *first* two coupling layers of a deeper flow. We agree that the visualizations of the deeper layers of the flows are less interpretable than the first layers due to the added noise, but you can still see the edges of the original inputs (or the outputs of the previous coupling layer) in the deep layers of the flows in Fig. 11 (e.g. you can clearly see the shape of the face in the s-activations for Celeb-A). • To test your hypothesis regarding the dominating contribution of the first layers in the likelihood, we visualize the contributions of the different layers to Jacobian log-determinant in Fig. 2. On average the contribution is relatively uniform across different layers while the first few layers show higher variance of the predicted scale s.

R4. Thank you for your review. We respectfully disagree with your assessment and we hope that you will consider raising your score. We would like to further clarify several misunderstandings. The primary focus of our paper is to provide empirical evidence for why flows assign high likelihoods to OOD data (Sections 5–7), and not to provide remedies to this issue. While we propose several approaches that improve the OOD detection results, we view them as understanding experiments that help us analyze how different design choices influence likelihood assignment on in-distribution and OOD data. • We do *not* argue that the poor performance of the flows comes solely from their normalization, MLE or learning from pixels. Instead, we argue that the OOD performance of a method is primarily decided by its inductive biases, and we argue that the inductive biases of the flows (analyzed in Sections 5–7) are not aligned with OOD detection (see Section 4 and Appendix A). • st-network capacity experiment: the smaller bottleneck size results in lower likelihood for train data, however, in the context of OOD detection we are interested in *relative ranking* of in-distribution versus OOD likelihoods. As mentioned above, this is not presented as a proposed solution but rather as a part of the analysis of flow's inductive biases.