We thank all reviewers for their critical comments and we address some questions below.

Q: (Reviewer 1) Ordering constraint on the multivariate quantile function
A: The reviewer is correct. Our definition of the quantile, similar to the SOS flow or other autoregressive models, assumes that we have fixed an ordering of coordinates. This is not an issue here since we perform novelty detection in the jointly learned latent space: choosing a different ordering simply amounts to permuting the latent space which will not affect the end result. Yes, other neural density estimators, such as autoregressive models (MAF, IAF, MADE, NAF, real NVP, or even the classic sigmoid belief networks), can also be plugged into our framework.

Q: (Reviewer 1) MGDA vs. two-stage training
A: Yes, we have tried the two-stage approach in our initial study. Similar as what many existing works reported [e.g. 1, 5, 39, 52], we found that it usually leads to suboptimal performance, possibly because the learned latent representation may not be necessarily helpful for training the second SOS stage. MGDA is more robust and parameter-free.

Q: (Reviewer 1) Statistical test on MNIST results
A: We did not perform statistical test since some baselines only used the default train/test split on MNIST. Calculating p-value based on 1 split does not seem to be meaningful. Nevertheless, we believe Table 3 suffices to show the competitiveness of our algorithms. Our setup here also comply with existing works [e.g. 1, 5, 16, 34, 38, 39, 51, 52].

Q: (Reviewer 2) Novelty and core contribution with regard to MQF
A: We apologize for the confusion, and we will follow the reviewer's suggestion to tone down our contribution in this definition and to adopt the more precise name "triangular quantile map" (TQM from now on).

In [9] Decurninge mentioned in passing that "For example, if we consider Rosenblatt transport," which is the only sentence that hinted the TQM definition. In Decurninge's dissertation, he expanded the discussion on the "Rosenblatt transport" which is essentially a constructive way to define (a version of) TQM. Decurninge's goal was to use TQM as an intermediate tool to define L-moments, rather than treating it as an object of independent interest. Moreover, our focus in this work, such as uniqueness, the importance of triangularity and monotonicity, an efficient estimation algorithm, and the application to novelty detection, was never touched in Decurninge's work. We will follow the reviewer's suggestion to tone down our role in this definition and discuss Decurninge's contribution in more details.

Thank you for bringing [Inouye and Ravikumar, 2018] to our attention! We will cite and discuss its relation to our work. The major difference is that we insist the mapping to be triangular and monotonic while Inouye and Ravikumar only require invertibility. Compared to Inouye and Ravikumar, TQM enjoys the following advantages: (a) Uniqueness. The density destructor is not unique hence unidentifiable. For instance, consider the trivial task of destructing the uniform density on $[0, 1]$: both $T(x) = x$ and $T(x) = 1 - x$ would do while the latter is not monotonic hence not allowed in our definition. (In high dimensions the unidentifiability issue is even worse and it would be difficult to define monotonicity in the absence of the triangular requirement.) (b) Computational efficiency. Inverting a monotonice triangular map only takes a linear number of 1-d root findings while inverting the destructor of Inouye and Ravikumar could be computationally challenging. Note that since we consider both the density and quantile rules, it is important to compute both the map Q and its inverse Q^{-1}. (c) Convenience. The triangular structure allows us to recycle many existing neural density estimators, such as the SOS flow and autoregressive models (MAF, IAF, NAF, real NVP, etc.).

Q: (Reviewer 2) When is the quantile approach useful?
A: We want to emphasize that we do not claim the quantile rule is better than the density rule, or vice versa. Our point is, through estimating the TQM, we now do not have to choose: both can be used at the same time. We have not tried to optimize the quantile rule either: the ones we used in the experiments were designed for simplicity and to encourage connectedness. On the other hand, thresholding the likelihood can create (many) highly disconnected (novel) regions, which is more flexible but also prune to outliers. It would be interesting to explore the tradeoff and to combine the two approaches in future work. We have run experiments on the donut example, see Figure 1. Note that if we change the quantile rule slightly to $\alpha < \|\mathbf{\Phi}(\mathbf{T}^{-1}(\tilde{\mathbf{z}})) - \frac{1}{2}\|_\infty < \beta$ with suitable α and β we can achieve similar performance as the density approach (which has an intrinsic advantage on this example due to disconnectedness of the novel regions). Note that since we perform novelty detection in the latent space $\mathbf{z}$, it is unlikely for the jointly learned latent space to be disconnected (as in the donut example). Finally, we point out that TQM allows one to visualize the progress of the estimation algorithm: the 1-d or 2-d projections of training samples onto coordinate axes should be uniform.

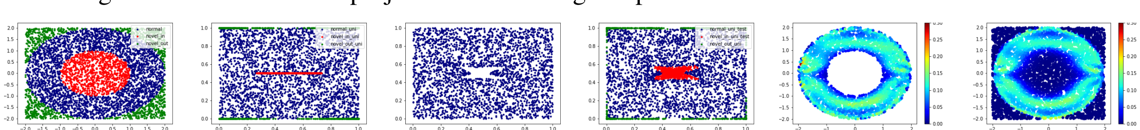

Figure 1: Donut example. From left to right: (1) uniform density over blue region; (2) pre-image in the cube using true TQM; (3) pre-image of training sample using estimated TQM; (4) pre-image of test sample using estimated TQM; (5) density of training sample using estimated TQM; (6) density of test sample using estimated TQM.