SUPPLEMENTARY FILE
Green Generative Modeling: Recycling Dirty Data using Recurrent Variational Autoencoders

Yu Wang * **Bin Dai** † **Gang Hua** ♯ **John Aston*** **David Wipf** ♯

* **University of Cambridge, Cambridge, UK**
† **Tsinghua University, Beijing, China**
♯**Microsoft Research, Beijing, China**
yw323@cam.ac.uk; daib13@mails.tsinghua.edu.cn; ganghua@microsoft.com; jada2@cam.ac.uk; davidwipf@gmail.com

1. Introduction

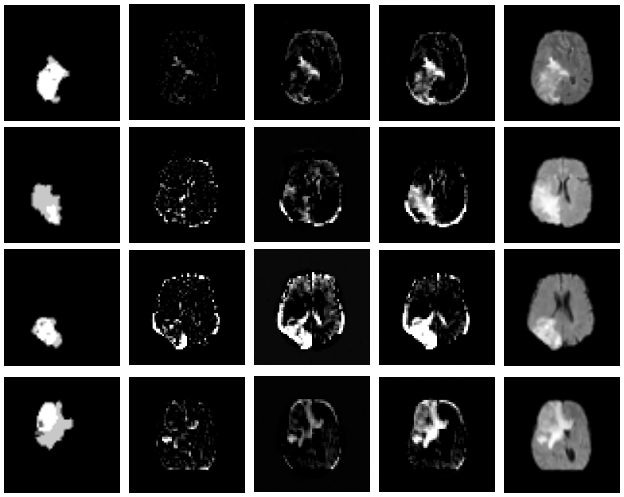
This document contains companion technical material regarding our UAI 2017 submission. Note that herein all equation numbers referencing back to the main submission document will be be prefixed with an 'M' to avoid confusion, i.e, (M.#) will refer to equation (#) from the main text. Similar notation differentiates sections, tables, and figures, e.g., Section M.#, etc. Notations without prefix or with the emphasized S.# will refer to the notations in this supplementary file.

2. Contents

This document includes the following supporting materials:

- **Section 3 - Medical Imaging Application:** We present additional results using the BRATS2015 brain tumor dataset (Menze et al., 2015), where the goal is the unsupervised detection of tumors. This provides preliminary evidence for the effectiveness of iC-VAE and RiC-VAE models on a challenging medical application.
- **Section 4 - Proof of Proposition M.1**
- **Section 5 - Alternative iC-VAE Derivation:** Here we re-derive the iC-VAE from the perspective of sparse Bayesian learning, which sheds additional insights into its robustness to outliers.
- **Section 6 - Deep Network Structure and Training Details:** This section specifies the DNN structures used for the RiC-VAE encoder and decoder layers referred to in Section M.5. We also discuss training details, e.g., learning rates, etc.
- **Section 7 - Additional Experimental Results:** Finally we describes additional supplementary experiments referred to in Section M.5 that could not fit in the main paper.

*. Y. Wang is sponsored by the EPSRC Centre for Mathematical Imaging in Healthcare, University of Cambridge, EP/N014588/1. B. Dai is finantially supported by Tsinghua University. Y. Wang and B. Dai are partially supported by sponsorship from Microsoft Research Asia.



(a) (b) (c) (d) (e)

Figure 1: Unsupervised brain tumor segmentation examples using BRATS2015 data. (a) Ground truth tumor segmentation annotation by human expert. (b) Abnormality segmentation by RPCA. (c) Abnormality segmentation by iC-VAE. (d) Abnormality segmentation by RiC-VAE(N=2,M=2). (e) Original brain image (preprocessed).

3. Medical Imaging Application: Unsupervised Tumor Detection

In this section, we examine various unsupervised learning algorithms with respect to their ability to detect abnormalities in real medical data. The public BRATS2015 data (Menze et al., 2015) consists of 274 multi-contrast MRI scans from brain glioma patients, out of which 54 were acquired from patients with low-grade tumors and 220 from patients with high-grade tumors. Each MRI scan contains a sequence of brain imaging slices consecutively taken from different brain depths. We collect the 80th-100th slices (hence some consistent depth cross different brains) from each flair channel of the high-grade MRI scans, and use these $21 \times 220 = 4620$ images to form training data.

We also resize all the original 240×240-resolution images to 60×60 (because of limited data for learning VAE networks), and normalize to zero mean and unit variance before application to unsupervised learning algorithms for evaluation purposes. The goal here is to compare the accuracy of detected tumor regions with respect to ground truth annotations labeled by human experts.

Table 1: Intersection over Union (BRATS 2015 dataset)

Algo.	Opt. Thr.	RPCA	iC-VAE	RiC-VAE
No D.O.	0.252	0.174	0.270	**0.349**
D.O	0.252	0.174	0.305	**0.352**

We train a iC-VAE for 5000 iterations (learning rate lr=1e-4) and use the resultant network weights to initialize the RiC-VAE(N=2,M=2) (lr=2e-5); the actual network structures used are described in Section 6 below. Both the iC-VAE and RiC-VAE were further trained for another 45000 iterations until convergence (in total 50000 iterations). We use IoU (Intersection over Union) to evaluate the tumor detection accuracy, which is defined as the ratio between the size of 'intersection of support sets' and the 'union of support sets' w.r.t. two sets: the ground truth tumor region and the predicted segmentation region.

In Table 1 we present IoU detection statistics for 4 different methods: (i) Opt. Thr., which denotes a prediction heuristic based on hard-thresholding the raw data set according to an optimal threshold (obtained by empirical search), (ii) Robust PCA (RPCA), a common unsupervised learning method for outlier removal (Candès et al., 2011), (iii) the iC-VAE described above, and (iv) the RiC-VAE(N=2,M=2) also from above. We also consider applying a dropout layer (D.O.) with rate 0.2 for both initialization purposes and final evaluation (this also affect the RiC-VAE via the initialization). Some complementary sample segmentation results are shown in Figure 1.

Both the evaluations in Figure 1 and Table 1 provide preliminary evidence of the effectiveness of iC-VAE and RiC-VAE(N=2,M=2) networks. By introducing the recycling mechanism via the RiC-VAE, multiple passes during training help to alleviate overfitting problems caused by the limited number of training data. Although extra structures and knowledge could be considered for overall performance improvement relative to supervised methods (e.g., to remove the ring artifacts around the brain borders, convolutional layers could be added to encoder and decoder networks), here our goal is merely proof-of-concept using unsupervised learning methods.

4. Proof of Proposition M.1

Using Jensen's inequality, we have that

$$\begin{aligned} 2\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}^{(i)})}\left[\log\left|x_j^{(i)}-\mu_{x_j}^{(i)}\right|\right] &\leq \log \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}^{(i)})}\left[\left(x_j^{(i)}-\mu_{x_j}^{(i)}\right)^2\right] \\ &= \inf_{\lambda_j^{(i)}>0} \frac{\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}^{(i)})}\left[\left(x_j^{(i)}-\mu_{x_j}^{(i)}\right)^2,\right]}{\lambda_j^{(i)}} + \log \lambda_j^{(i)} \end{aligned} \tag{1}$$

where an irrelevant constant has been omitted and $\boldsymbol{\lambda}^{(i)} \in \mathbb{R}_+^d$ for all i represent arbitrary variational parameters, independent of $\boldsymbol{z}$. Any concave function can be expressed as a minimization of upper-bounding linear functions in this way (Jordan et al., 1999). The expectation now admits a closed-form solution leading to

$$\begin{aligned} &\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}^{(i)})}\left[\left(\boldsymbol{x}^{(i)}-\boldsymbol{W}\boldsymbol{z}-\boldsymbol{b}\right)^\top\left(\boldsymbol{\Lambda}^{(i)}\right)^{-1}\left(\boldsymbol{x}^{(i)}-\boldsymbol{W}\boldsymbol{z}-\boldsymbol{b}\right)\right] \\ &= \left(\boldsymbol{x}^{(i)}-\boldsymbol{W}\boldsymbol{\mu}_z^{(i)}-\boldsymbol{b}\right)^\top\left(\boldsymbol{\Lambda}^{(i)}\right)^{-1}\left(\boldsymbol{x}^{(i)}-\boldsymbol{W}\boldsymbol{\mu}_z^{(i)}-\boldsymbol{b}\right)+\operatorname{tr}\left[\boldsymbol{\Sigma}_z^{(i)}\boldsymbol{W}^\top\left(\boldsymbol{\Lambda}^{(i)}\right)^{-1}\boldsymbol{W}\right]. \end{aligned} \tag{2}$$

where $\boldsymbol{\Lambda}^{(i)} = \text{diag}[\boldsymbol{\lambda}^{(i)}]$. Using these expressions we obtain the new upper bound on the original VAE cost given by

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{X}) \;\leq\; & \sum_i \inf_{\boldsymbol{\Lambda}^{(i)} \succ \mathbf{0}} \Big\{ \text{tr}\left[\boldsymbol{\Sigma}_z^{(i)}\right] - \log\left|\boldsymbol{\Sigma}_z^{(i)}\right| + \|\boldsymbol{\mu}_z^{(i)}\|_2^2 \\ & + \left(\boldsymbol{x}^{(i)} - \boldsymbol{W}\boldsymbol{\mu}_z^{(i)} - \boldsymbol{b}\right)^\top \left(\boldsymbol{\Lambda}^{(i)}\right)^{-1} \left(\boldsymbol{x}^{(i)} - \boldsymbol{W}\boldsymbol{\mu}_z^{(i)} - \boldsymbol{b}\right) \\ & + \text{tr}\left[\boldsymbol{\Sigma}_z^{(i)} \boldsymbol{W}^\top \left(\boldsymbol{\Lambda}^{(i)}\right)^{-1} \boldsymbol{W}\right] + \log\left|\boldsymbol{\Lambda}^{(i)}\right| \Big\}. \end{aligned} \tag{3}$$

Because $\boldsymbol{\mu}_z^{(i)}$ and $\boldsymbol{\Sigma}_z^{(i)}$ appear in different terms, we can optimize over each separately in terms of $\boldsymbol{\Lambda}^{(i)}$ and $\boldsymbol{W}$ by taking gradients, equating to zero, and solving. This leads to the optimal solutions

$$\begin{aligned}\boldsymbol{\mu}_z^{(i)} &= \boldsymbol{W}^\top \left(\boldsymbol{\Lambda}^{(i)} + \boldsymbol{W}\boldsymbol{W}^\top\right)^{-1} \left(\boldsymbol{x}^{(i)} - \boldsymbol{b}\right), \\ \boldsymbol{\Sigma}_z^{(i)} &= \left[\boldsymbol{W}^\top \boldsymbol{\Lambda}^{(i)} \boldsymbol{W} + \boldsymbol{I}\right]^{-1}. \end{aligned} \tag{4}$$

Plugging these values into (3) and ignoring constants, we obtain the stated bound. Moreover, we observe from this development that in fact $\boldsymbol{\mu}_z^{(i)}$ and $\boldsymbol{\Sigma}_z^{(i)}$ need not involve overly complex deep network structures. By satisfying (4), which can be viewed as simple affine stationary conditions, the result still holds.

5. Alternative iC-VAE Derivation

In this section, we re-derive the iC-VAE objective (M.8) from the perspective of sparse Bayesian learning or automatic relevance determination applied to low-rank modeling (Wipf, 2012).

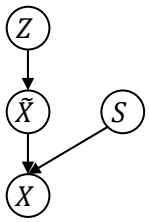


Figure 2: Graphical Model

We begin with the graphical model from Figure 2. We have a latent variable $\boldsymbol{z}$ which determines the 'clean' data $\tilde{\boldsymbol{x}}$, which is then corrupted by some sparse noise $\boldsymbol{s}$, which follows the Jeffreys prior $p(\boldsymbol{s}) \propto \Pi_{j=1}^d \frac{1}{|s_j|}$. Then we obtain the observed data $\boldsymbol{x}$, and $p(\boldsymbol{x})$ can be expressed as

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{s}) p(\boldsymbol{z}, \boldsymbol{s}) d\boldsymbol{z} d\boldsymbol{s} \tag{5}$$

Similarly, we can write a lower bound on the negative log of $p(\boldsymbol{x})$ by replacing the latent $\boldsymbol{z}$ in a conventional VAE with $(\boldsymbol{z}, \boldsymbol{s})$, giving us

$$\mathcal{L} = \mathbb{KL}\left[q(\boldsymbol{z}, \boldsymbol{s}|\boldsymbol{x})||p(\boldsymbol{z}, \boldsymbol{s})\right] + \mathbb{E}_{(\boldsymbol{z},\boldsymbol{s})\sim q(\boldsymbol{z},\boldsymbol{s}|\boldsymbol{x})}\left[-\log p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{s})\right]. \tag{6}$$

Note that we omit θ and ϕ in the loss function because they serve different roles in this formulation (in fact they are both absorbed into q). We define the prior of $\boldsymbol{z}$ as a standardized Gaussian distribution and the approximated posterior of $\boldsymbol{z}$ as a Gaussian distribution determined by the encoder. The prior of $\boldsymbol{s}$ is defined as the Jeffreys prior as mentioned before while the approximated posterior of $\boldsymbol{s}$ is a Dirac-delta function. In summary then we have

$$\begin{aligned}
q(\boldsymbol{z}, \boldsymbol{s}|\boldsymbol{x}) &= q(\boldsymbol{z}|\boldsymbol{x})q(\boldsymbol{s}|\boldsymbol{z}, \boldsymbol{x}) \\
q(\boldsymbol{z}|\boldsymbol{x}) &\sim \mathcal{N}\left(\boldsymbol{\mu}_z(\boldsymbol{x}), \boldsymbol{\Sigma}_z(\boldsymbol{x})\right) \\
q(\boldsymbol{s}|\boldsymbol{z}, \boldsymbol{x}) &\sim \delta(\boldsymbol{x} - \boldsymbol{\mu}_x(\boldsymbol{z})) \\
p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{s}) &\propto \exp\left(-\frac{1}{2\lambda}||\boldsymbol{x} - \boldsymbol{\mu}_x(\boldsymbol{z}) - \boldsymbol{s}||^2_{\mathcal{F}}\right),
\end{aligned} \tag{7}$$

where $\delta(\cdot)$ is a Dirac-delta function. Now consider the second term on the r.h.s of (6). It becomes

$$\begin{aligned}
&\mathbb{E}_{(\boldsymbol{z},\boldsymbol{s})\sim q(\boldsymbol{z},\boldsymbol{s}|\boldsymbol{x})}\left[\frac{1}{2\lambda}||\boldsymbol{x} - \boldsymbol{\mu}_x - \boldsymbol{s}||^2_{\mathcal{F}} + C\right] \\
=&\mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}|\boldsymbol{x})}\left[\int \delta(\boldsymbol{x} - \boldsymbol{\mu}_x)\frac{1}{2\lambda}||\boldsymbol{x} - \boldsymbol{\mu}_x - \boldsymbol{s}||^2_{\mathcal{F}} d\boldsymbol{s}\right] + C \\
=&\mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}|\boldsymbol{x})}\left[\frac{1}{2\lambda}||\boldsymbol{x} - \boldsymbol{\mu}_x - (\boldsymbol{x} - \boldsymbol{\mu}_x)||^2_{\mathcal{F}}\right] + C \\
=&C,
\end{aligned} \tag{8}$$

where C is a constant, hence we can omit this term. The first term of (6) can be decomposed into two parts as

$$\begin{aligned}
&\int q(\boldsymbol{z}|\boldsymbol{x})q(\boldsymbol{s}|\boldsymbol{z}, \boldsymbol{x}) \log \frac{q(\boldsymbol{z}|\boldsymbol{x})q(\boldsymbol{s}|\boldsymbol{z}, \boldsymbol{x})}{p(\boldsymbol{z})p(\boldsymbol{s})} d\boldsymbol{z}d\boldsymbol{s} \\
=&\int q(\boldsymbol{z}|\boldsymbol{x}) \log \frac{q(\boldsymbol{z}|\boldsymbol{x})}{p(\boldsymbol{z})} d\boldsymbol{z} + \int q(\boldsymbol{z}|\boldsymbol{x})q(\boldsymbol{s}|\boldsymbol{z}, \boldsymbol{x}) \log \frac{q(\boldsymbol{s}|\boldsymbol{z}, \boldsymbol{x})}{p(\boldsymbol{s})} d\boldsymbol{z}d\boldsymbol{s} \\
=&\mathbb{KL}\left[q(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\right] + \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}|\boldsymbol{x})}\left[\int q(\boldsymbol{s}|\boldsymbol{z}, \boldsymbol{x}) \log \frac{q(\boldsymbol{s}|\boldsymbol{z}, \boldsymbol{x})}{p(\boldsymbol{s})} d\boldsymbol{s}\right].
\end{aligned} \tag{9}$$

The first term is exactly the same as the KL term in the conventional VAE. Plugging $q(\boldsymbol{s}|\boldsymbol{z},\boldsymbol{x}) = \delta(\boldsymbol{s}; \boldsymbol{x} - \boldsymbol{\mu}_x(\boldsymbol{z}))$ into the second term gives

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}|\boldsymbol{x})}\left[\int q(\boldsymbol{s}|\boldsymbol{z},\boldsymbol{x}) \log \frac{q(\boldsymbol{s}|\boldsymbol{z},\boldsymbol{x})}{p(\boldsymbol{s})} d\boldsymbol{s}\right] \\
=&\mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}|\boldsymbol{x})}\left[\int \delta(\boldsymbol{s}; \boldsymbol{x} - \boldsymbol{\mu}_x) \log \delta(\boldsymbol{s}; \boldsymbol{x} - \boldsymbol{\mu}_x) d\boldsymbol{s} - \int \delta(\boldsymbol{s}; \boldsymbol{x} - \boldsymbol{\mu}_x) \log p(\boldsymbol{s}) d\boldsymbol{s}\right] \\
=&C - \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}|\boldsymbol{x})}\left[\log p(\boldsymbol{s} = \boldsymbol{x} - \boldsymbol{\mu}_x)\right] \\
=&C + \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}|\boldsymbol{x})}\left[\sum_j \log \left|x_j - \mu_{x_j}\right|\right].
\end{aligned} \tag{10}$$

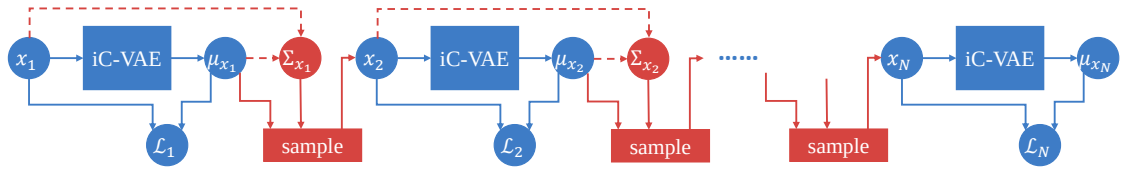
It should be noticed that though $\int \delta(\boldsymbol{s}; \boldsymbol{x} - \boldsymbol{\mu}_x) \log \delta(\boldsymbol{s}; \boldsymbol{x} - \boldsymbol{\mu}_x) d\boldsymbol{s}$ is infinite, it is not related to the parameters thus it can be treated as a constant. Combining these results, the lower bound of the negative log likelihood can be finally be expressed as

$$\mathcal{L} = \mathbb{KL}[q(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})] + \sum_j \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \left|x_j - \mu_{x_j}\right|\right] + C, \tag{11}$$

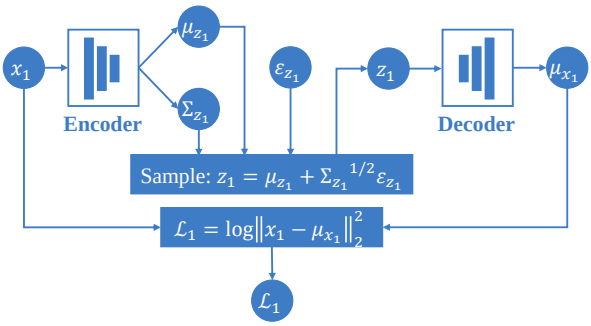
which is equivalent to (M.8) in the main paper once we account for multiple samples $i = 1, \ldots, n$. This formulation explicitly elucidates the crucial role that outliers play in the iC-VAE development.

6. Deep Network Structure and Training Details

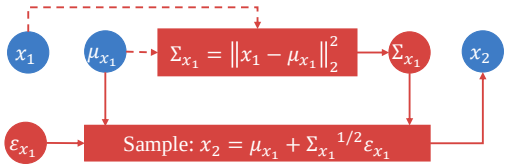
In this section, we illustrate the further particulars of the RiC-VAE framework originally shown in Figure M.1. Here Figure 3 first presents the general structure flow, including additional ingredients pertinent to the iC-VAE.



(a) Pipeline of Recurrent Variational Autoencoder



(b) Details of iC-VAE



(c) Details of Sample Block

Figure 3: Detailed structure flow of the RiC-VAE network.

6.1 RiC-VAE network structure

The dimensions of each input layer, encoder inner-product hidden layers, latent dimension of $\boldsymbol{z}$, decoder inner-product hidden layers, and finally each output layer are listed here for each data set.

Frey face (Section M.5.2)
$560(input) - 1000 - 500 - 250 - 10(\boldsymbol{z}) - 250 - 500 - 1000 - 560(output)$
We use ReLU activations for all inner product layers except the last one which uses sigmoid activations to accommodate the magnitude range of the image pixels.

MNIST (Section M.5.3.)
$784(input) - 1000 - 500 - 250 - 30(\boldsymbol{z}) - 250 - 500 - 1000 - 784(output)$.
The activation scheme is the same as for the Frey face data.

BRATS2015 (Section 3)
$3600(input) - 50 - 10(\boldsymbol{z}) - 50 - 3600(output)$
The activation scheme is also the same as Frey face data except the last layer, after which we use no activation. This is because the format of BRATS2015 is different from ordinary images.

6.2 Training Details

Learning rates and iteration counts are listed below.

Frey face
For iC-VAE, we set the learning rate as 1×10^{-4} and train 150000 iterations with batch size equal to 100. For RiC-VAE(N=2), we use the same training setting. Since the samples are actually doubled, we halve the learning rate to be 5×10^{-5}.

MNIST
For iC-VAE, we set the learning rate as 1×10^{-4} and train 600000 iterations with batch size equal to 100. Then we use this to initialize the RiC-VAE(N=5) and train an extra 100000 iterations with learning rate 2×10^{-5}. To compare iC-VAE and RiC-VAE fairly, we also train the iC-VAE for another 100000 iterations.

BRATS2015
For the iC-VAE, we set the learning rate as 1×10^{-4} and train 5000 iterations with batch size equal to 100. Then we use this to initialize the RiC-VAE(N=2) and train an extra 45000 iterations with learning rate 5×10^{-5}. We also train the iC-VAE for another 45000 iterations for fair comparison.

7. Additional Experimental Results

This section contains a wider range of empirical results that complement those from Section M.5 in the main paper.

7.1 Frey face recovery from new outlier patterns

When testing the models on Frey face dataset (Section M.5.2), we synthesize new test images by changing the corruption patterns seen during training. Instead of using circle outliers as in the training data, new 'rectangle' dirty patterns having random width, length and location were added. In this way, we examine the resilience of models on the outlier distribution distinct from the training data. Figure 4 demonstrates the dirty pattern in test mode and the corresponding reconstruction results when trained at a 60% corruption level.

Figure 4: Visualizing reconstructions on Frey face test data. *Row 1*: Dirty test data. *Row 2:* Reconstruction by iC-VAE. *Row 3:* Reconstruction by RiC-VAE (N=2,M=2). *Row 4:* Original clean Frey faces

7.2 Generative performance using MNIST data

This section provides supporting material for Section M.5.3. For illustration purposes, Figure 5 first presents some random samples from both the original clean MNIST data and the corrupted version at a 40% noise level. Later Figures 6-10 visualize samples generated by VAE networks under various clean and dirty scenarios. Among other things, these figures illustrate that once trained, the RiC-VAE can produce reasonable samples after cycling through the network a variable number of times.

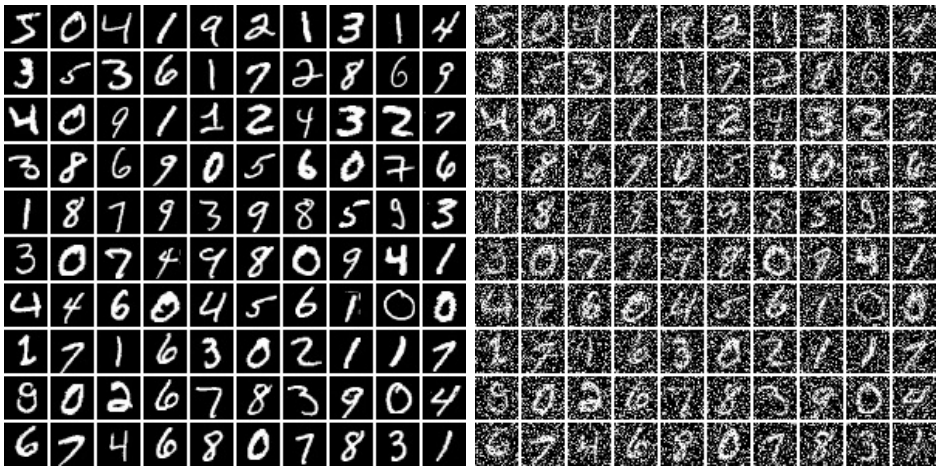
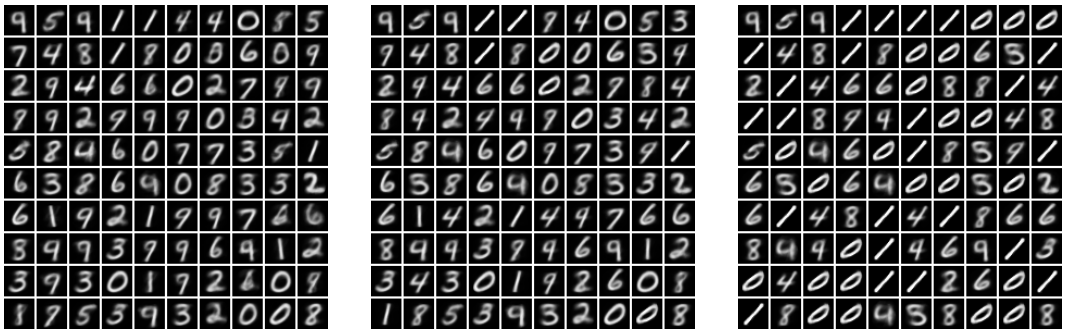
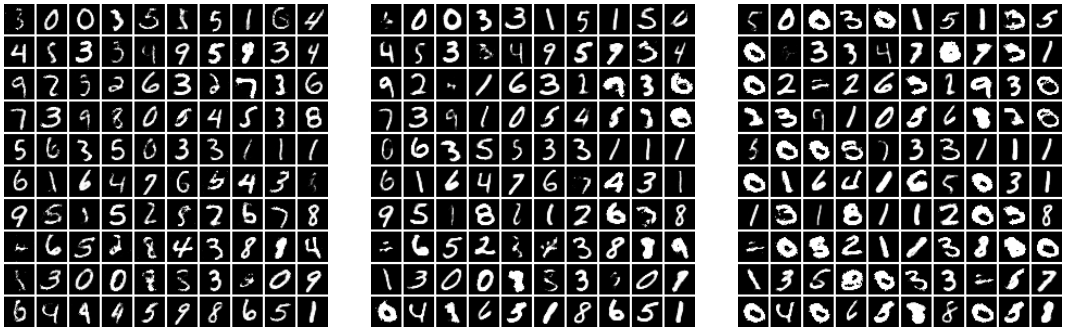


Figure 5: *Left*: Original clean MNIST data. *Right*: MNIST training data corrupted with 40% salt-and-pepper noise.



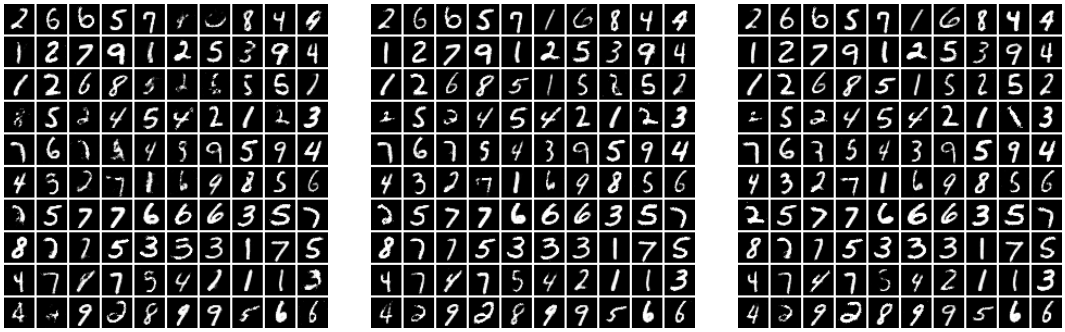
(a) (b) (c)

Figure 6: Samples generated from ℓ_2-VAE and Rℓ_2-VAE models trained on clean MNIST data. (a) ℓ_2-VAE, (b) Rℓ_2-VAE(N=1,M=5), and (c) Rℓ_2-VAE(N=1,M=20).



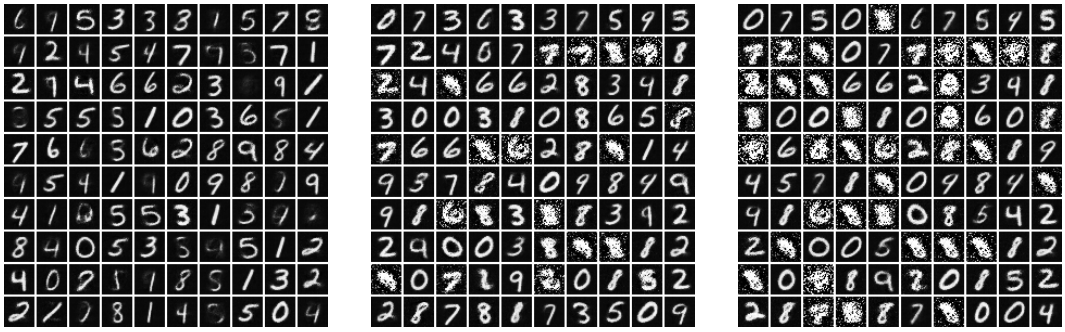
(a) (b) (c)

Figure 7: Samples generated from iC-VAE and RiC-VAE models trained on clean MNIST data. (a) iC-VAE, (b) RiC-VAE(N=1,M=5), and (c) RiC-VAE(N=1,M=20).



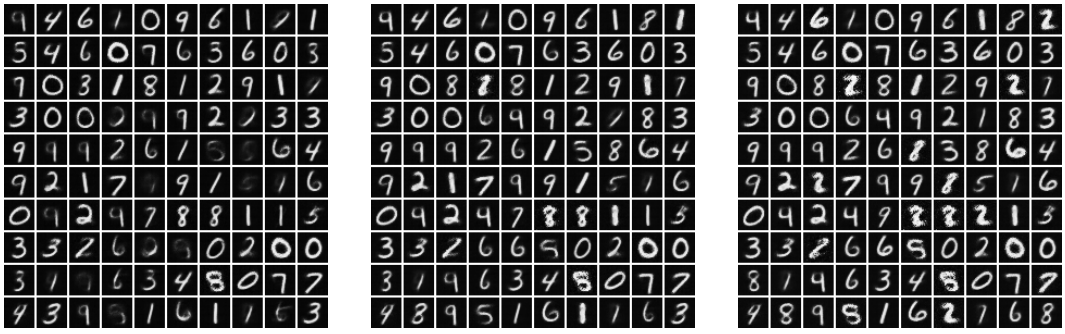
(a) (b) (c)

Figure 8: Samples generated from RiC-VAE(N=5,M=#) having different M and trained on clean MNIST data. (a) RiC-VAE(N=5,M=1), (b) RiC-VAE(N=5,M=5), and (c) RiC-VAE(N=5,M=20).



(a) (b) (c)

Figure 9: Samples generated from iC-VAE and RiC-VAE trained on noisy MNIST data. (a) iC-VAE, (b) RiC-VAE(N=1,M=5), and (c) RiC-VAE (N=1,M=20).



(a) (b) (c)

Figure 10: Samples generated from the dirty training set using RiC-VAE(N=5,M=#) and different M. (a) RiC-VAE(N=5,M=1), (b) RiC-VAE(N=5,M=5), and (c) RiC-VAE(N=5,M=20).

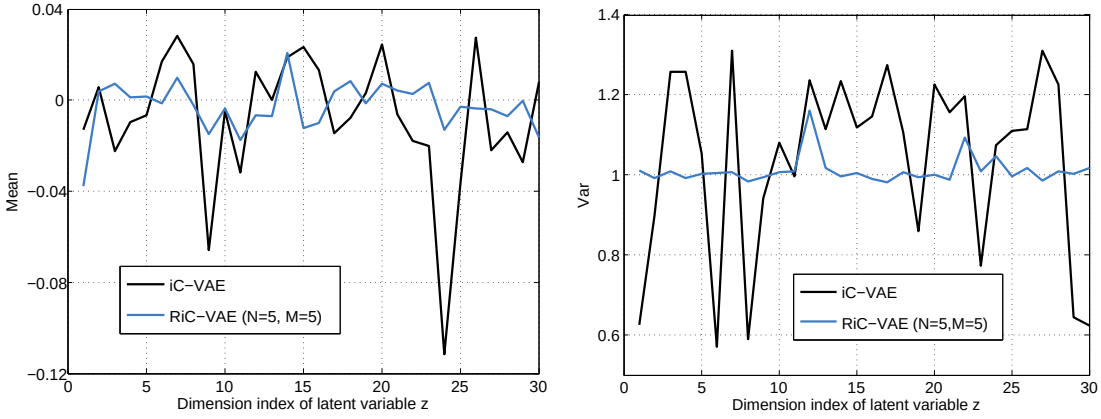
7.3 Further statistical validation of generated samples

The approximate marginal distribution of the latent variable $\boldsymbol{z}$ should satisfy

$$\frac{1}{n}\sum_i q_\phi\left(\boldsymbol{z}|\boldsymbol{x}^{(i)}\right) \approx p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z};\boldsymbol{0},\boldsymbol{I}) \tag{12}$$

if our estimated distributions are accurate. In the main paper we visualized 2D samples from this distribution to test this approximation. Here we consider a higher dimensional comparisons with a standardized Gaussian.

In particular, we draw 60000 samples from $\frac{1}{n}\sum_i q_\phi\left(\boldsymbol{z}|\boldsymbol{x}^{(i)}\right)$ and then compute the mean and variance of each of the $\kappa = 30$ dimensions in $\boldsymbol{z}$ as used with the MNIST data. Ideally these means should be near zero while the variances should be near one. Figure 11 shows the results for an iC-VAE and a RiC-VAE. From this figure it is immediately apparent that the RiC-VAE statistics are far more consistent with the ideal standardized Gaussian (which emerges if the non-Gaussian underlying latent distributions are approximated well) than the iC-VAE, suggesting that our recycling approach indeed better captures the true underlying distributions. Note also that results are similar as M is varied, implying that the statistics of recycled data or multiple passes at test time are stable.



(a) Mean value comparison.

(b) Variance value comparison

Figure 11: Means and variances of samples drawn from $\frac{1}{n}\sum_i q_\phi\left(\boldsymbol{z}|\boldsymbol{x}^{(i)}\right)$.

References

Candès E. J., Li X., Ma Y., et al. Robust principal component analysis?. *Journal of the ACM (JACM)*, 58(3), 2011.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., Saul, L. K.. An introduction to variational methods for graphical models *Machine learning*, 37(2): 1999

Menze B H, Jakab A, Bauer S, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE transactions on medical imaging*,34(10): 2015

Wipf D. Non-convex rank minimization via an empirical Bayesian approach. *Uncertainty in Artificial Intelligence, 2012.*