

## A APPENDIX

### A.1 OBJECTIVE OF VAES

The objective of VAES is to maximize a tractable variational lower bound on the data log-likelihood, called the Evidence Lower Bound (ELBO):

$$\mathbb{E}_{p(\mathbf{x}_e)} [\mathbb{E}_{p(\mathbf{z}_e|\mathbf{x}_e)} [\log p(\mathbf{x}_d|\mathbf{z})] - \mathcal{D}_{\text{KL}}(p(\mathbf{z}_e|\mathbf{x}_e) || p(\mathbf{z}_d))] . \quad (5)$$

It can be also shown that the objective of VAES is equivalent to minimizing the KL divergence (or maximizing the negative KL divergence) between the encoding and the decoding distributions (Livne et al., 2019; Esmaili et al., 2019; Pu et al., 2017b; Chen et al., 2018):

$$-\mathcal{D}_{\text{KL}}(p(\mathbf{x}_e, \mathbf{z}_e) || p(\mathbf{x}_d, \mathbf{z}_d)) = \mathbb{E}_{p(\mathbf{x}_e, \mathbf{z}_e)} \left[ \log \frac{p(\mathbf{x}_d, \mathbf{z}_d)}{p(\mathbf{z}_e|\mathbf{x}_e)} \right] - \mathbb{E}_{p(\mathbf{x}_e)} [\log p(\mathbf{x}_e)] . \quad (6)$$

The right hand side of equation 6 is only different from equation 5 in terms of a constant, which is the entropy of the observed data.

### A.2 PROOF OF THEOREM 1

The proof extends that of Theorem 1 in (Tolstikhin et al., 2018). In particular, (Tolstikhin et al., 2018) aims to minimize the OT cost of the marginal distributions  $p(\mathbf{x}_e)$  and  $p(\mathbf{x}_d)$ , and the proof there is based on the joint probability of three random variables: the observed data, the generated data, and the latent representation. In contrast, we propose to minimize the OT cost of the joint distributions of the observed data and the latent representation induced by the encoder and the decoder. As a result our proof is based on the joint distribution of four random variables  $(\mathbf{x}_e, \mathbf{z}_e, \mathbf{x}_d, \mathbf{z}_d) \in \mathcal{X} \times \mathcal{Z} \times \mathcal{X} \times \mathcal{Z}$ . We assume that the joint distribution  $p(\mathbf{x}_e, \mathbf{z}_e, \mathbf{x}_d, \mathbf{z}_d)$  satisfies the following three conditions:

1.  $\mathbf{e} \triangleq (\mathbf{x}_e, \mathbf{z}_e) \sim p(\mathbf{x}_e)p(\mathbf{z}_e|\mathbf{x}_e)$ ;
2.  $\mathbf{d} \triangleq (\mathbf{x}_d, \mathbf{z}_d) \sim p(\mathbf{z}_d)p(\mathbf{x}_d|\mathbf{z}_d)$ ; and
3.  $\mathbf{x}_d \perp\!\!\!\perp \mathbf{x}_e | \mathbf{z}_d$  (conditional independence).

The first two conditions specify the encoder and the decoder respectively, and the last condition indicates that given the latent prior the generated data and the observed data are independent.

Denote the set of the above joint distributions as  $\mathcal{P}(\mathbf{x}_e, \mathbf{z}_e, \mathbf{x}_d, \mathbf{z}_d)$ . Obviously, we have  $\mathcal{P}(\mathbf{x}_e, \mathbf{z}_e, \mathbf{x}_d, \mathbf{z}_d) \subseteq \mathcal{P}(\mathbf{e} \sim p(\mathbf{e}), \mathbf{d} \sim p(\mathbf{d}))$  due to the third condition. If the decoder is deterministic,  $p(\mathbf{x}_d|\mathbf{z}_d)$  is a Dirac distribution thus  $\mathcal{P}(\mathbf{x}_e, \mathbf{z}_e, \mathbf{x}_d, \mathbf{z}_d) = \mathcal{P}(\mathbf{e} \sim p(\mathbf{e}), \mathbf{d} \sim p(\mathbf{d}))$ . With this result, we can rewrite the objective of the underlying OT problem as follows:

$$\begin{aligned} W_c(p(\mathbf{e}), p(\mathbf{d})) &= \inf_{\Gamma \in \mathcal{P}(\mathbf{x}_e, \mathbf{z}_e, \mathbf{x}_d, \mathbf{z}_d)} \mathbb{E}_{(\mathbf{e}, \mathbf{d}) \sim \Gamma} c(\mathbf{e}, \mathbf{d}) \\ &= \inf_{\Gamma \in \mathcal{P}(\mathbf{x}_e, \mathbf{z}_e, \mathbf{z}_d)} \mathbb{E}_{(\mathbf{x}_e, \mathbf{z}_e, \mathbf{z}_d) \sim \Gamma} c(\mathbf{e}, \mathbf{d}) \end{aligned} \quad (7)$$

$$= \inf_{p(\mathbf{z}_e|\mathbf{x}_e), p(\mathbf{z}_d|\mathbf{x}_e, \mathbf{z}_e)} \mathbb{E}_{p(\mathbf{x}_e)} \mathbb{E}_{p(\mathbf{z}_e|\mathbf{x}_e)} \mathbb{E}_{p(\mathbf{z}_d|\mathbf{x}_e, \mathbf{z}_e)} c(\mathbf{e}, \mathbf{d}) \quad (8)$$

$$= \inf_{p(\mathbf{z}_d|\mathbf{x}_e)} \mathbb{E}_{p(\mathbf{x}_e)} \mathbb{E}_{p(\mathbf{z}_d|\mathbf{x}_e)} c(\mathbf{e}, \mathbf{d}), \quad (9)$$

where in equation 7  $\mathcal{P}(\mathbf{x}_e, \mathbf{z}_e, \mathbf{z}_d)$  denotes the set of the joint distributions of  $(\mathbf{x}_e, \mathbf{z}_e, \mathbf{z}_d)$  induced by  $\mathcal{P}(\mathbf{x}_e, \mathbf{z}_e, \mathbf{x}_d, \mathbf{z}_d)$  and it holds due to the deterministic decoder, and equation 9 holds due to the deterministic encoder.

### A.3 COMPARISON BETWEEN SWAES ( $\beta = 1$ ) WITH WAES

The objective of SWAES ( $\beta = 1$ ) minimizes the OT cost between the joint distributions of the data and the latent, *i.e.*,  $W_c(p(\mathbf{e}), p(\mathbf{d}))$ , while the objective of WAES (Tolstikhin et al., 2018) minimizes the OT cost between the marginal distributions of the data, *i.e.*,  $W_c(p(\mathbf{x}_e), p(\mathbf{x}_d))$ , where  $p(\mathbf{x}_d)$  is the marginal data distribution induced by the decoding distribution  $p(\mathbf{d})$ . The problem of WAES is

first formulated as an optimization with the constraint  $p(\mathbf{z}_e) = p(\mathbf{z}_d)$ , where  $p(\mathbf{z}_e)$  is the marginal distribution induced by the encoding distribution  $p(\mathbf{e})$ , and then relaxed by adding a regularizer. With the deterministic decoder, the final optimization problem of WAEs is as follows:

$$\inf_{p(\mathbf{z}_e|\mathbf{x}_e)} \mathbb{E}_{p(\mathbf{x}_e)} \mathbb{E}_{p(\mathbf{z}_e|\mathbf{x}_e)} c(\mathbf{x}_e, D(\mathbf{z}_e)) + \lambda \mathcal{D}(p(\mathbf{z}_e), p(\mathbf{z}_d)), \quad (10)$$

where  $\mathcal{D}(\cdot)$  denotes some divergence measure. Comparing equation 10 to equation 3, we can see that both methods decompose the loss into the losses in the data and the latent spaces. Differently, in equation 10 the first term reflects the reconstruction loss in the data space and the second term represents the distribution-based dissimilarity in the latent space; while in equation 3 the  $\mathbf{x}$ -loss is closely related to the denoising and the generation quality and the  $\mathbf{z}$ -loss measures the sample-based dissimilarity. Moreover, equation 10 is optimized over the posterior  $p(\mathbf{z}_e|\mathbf{x}_e)$  with a fixed prior  $p(\mathbf{z}_d)$ , while equation 3 is optimized over the conditional prior  $p(\mathbf{z}_d|\mathbf{x}_e)$  with a potentially learnable prior.

#### A.4 DATASETS AND NETWORK ARCHITECTURES

In this section, we briefly describe the datasets, the network architectures, and the hyperparameters that are used in our training algorithm.

- **MNIST:** The dataset includes 70,000 gray-level images of numeric digits from 0 to 9, each of the size  $28 \times 28$ . There are 7,000 images per class. The training set contains 50,000 images, the validation set contains 10,000 images for choosing the best model based on the loss function, and the test set contains 10,000 images.
- **Fashion-MNIST:** The dataset includes 70,000 gray-level images of fashion products in 10 classes. This dataset has the same image size and the split of the training, validation, and test sets as in MNIST.
- **Coil20:** The dataset includes gray-scale images of 20 objects, each image of the size  $32 \times 32$ . The training set contains 1040 images, the validation set contains 200 images for choosing the best model based on the loss function, and the test set contains 200 images.
- **CIFAR10-sub:** The CIFAR-10 dataset consists of 60,000  $32 \times 32$  colour images in 10 classes with 6,000 images per class. There are 40,000 training, 10,000 validation, and 10,000 test images. We randomly select three classes to form the CIFAR10-sub dataset, namely *bird*, *cat*, and *ship*.

**Network architecture of SWAE:** The building block of the network structure of SWAE is based on VampPrior, called GatedConv2d. GatedConv2d contains two convolutional layers with the gating mechanism utilized as an element-wise non-linearity. The parameters in the function GatedConv2d() represent the number of the input channels, the number of the output channels, kernel size, stride, and padding, respectively. The conditional prior network outputs the mean and the log-variance of a Gaussian distribution, based on which the latent prior is sampled.

- The structure of the encoder network: GatedConv2d(1,32,7,1,3)-GatedConv2d(32,32,3,2,1)-GatedConv2d(32,64,5,1,2)-GatedConv2d(64,64,3,2,1)-GatedConv2d(64,6,3,1,1), followed by one fully-connected layer with no activation function.
- The structure of the conditional prior network: The layers of GatedConv2d are the same as those in the encoder network, which are followed by two fully-connected layers. One produces the mean, and the other produces the log-variance with the activation function Hardtanh.
- The structure of the decoder network: Two fully-connected layers with the gating mechanism, followed by GatedConv2d(1,64,3,1,1)-GatedConv2d(64,64,3,1,1)-GatedConv2d(64,64,3,1,1)-GatedConv2d(64,64,3,1,1), followed by a convolutional layer with the activation function Sigmoid.

The algorithm is trained by Adam with the learning rate  $= 0.001$ ,  $\beta_1 = 0.9$ , and  $\beta_2 = 0.999$ .

**Setup of the number of the pseudo-inputs  $K$ :** As suggested in (Tomczak & Welling, 2018; Livne et al., 2019) we set the value of  $K$  in VampPrior and MIM on MNIST and Fashion-MNIST to 500.

We found  $K = 500$  is also suitable for VampPrior and MIM on Coil20 and CIFAR10-sub. Unlike VampPrior and MIM, for SWAE we found that increasing  $K$  improves the performance and we set  $K$  to 4000 on MNIST, Fashion-MNIST, and CIFAR10-sub. Coil20 is a relatively small dataset and we set  $K$  to 500 for SWAE, VampPrior, and MIM.

#### A.5 MORE EXPERIMENTAL RESULTS

In this section, we show more experimental results based on the comparison with the benchmarks.

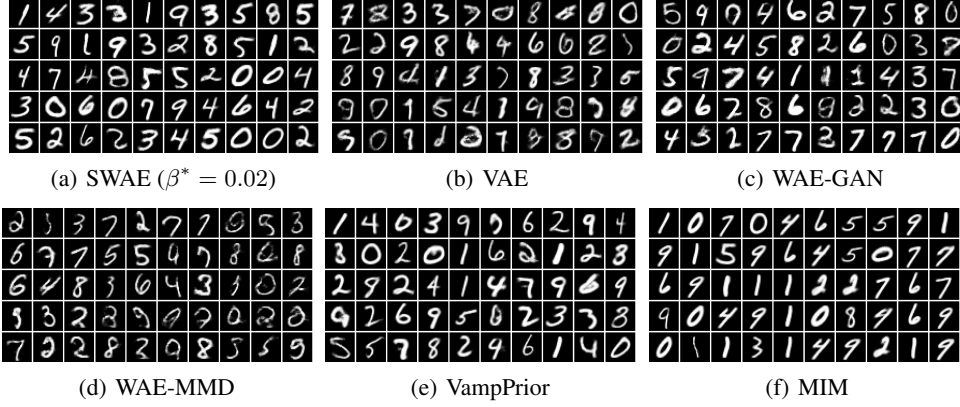


Figure 5: Generated new samples on MNIST.  $\dim\mathbf{z} = 8$  for all methods.

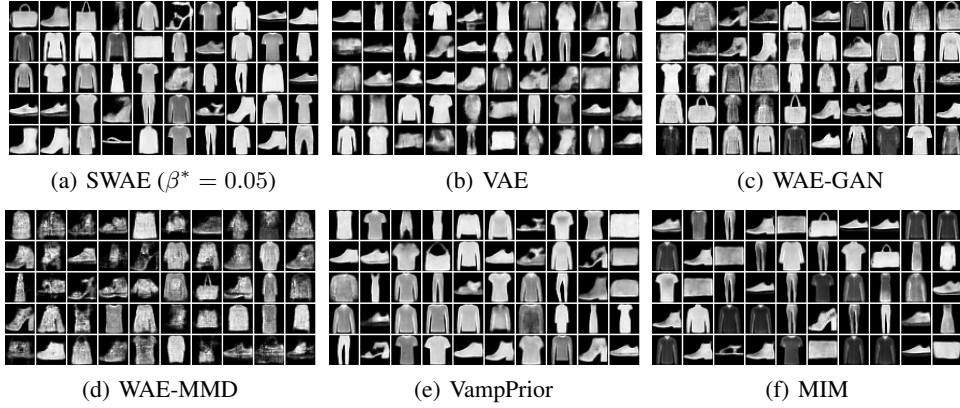
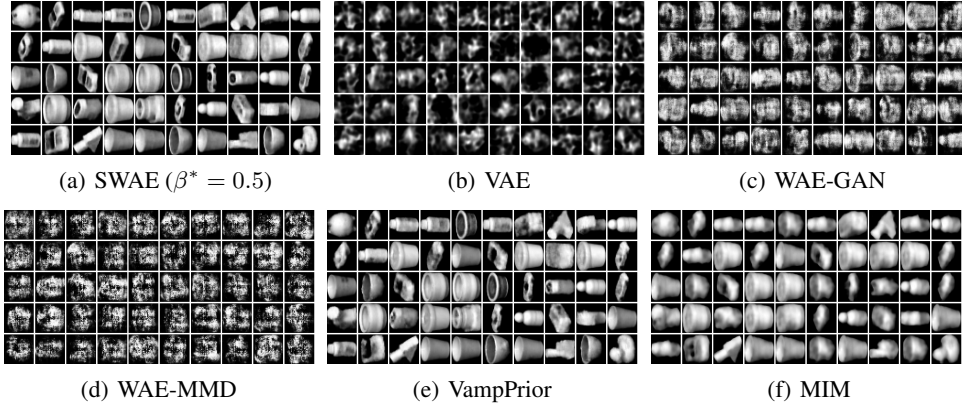
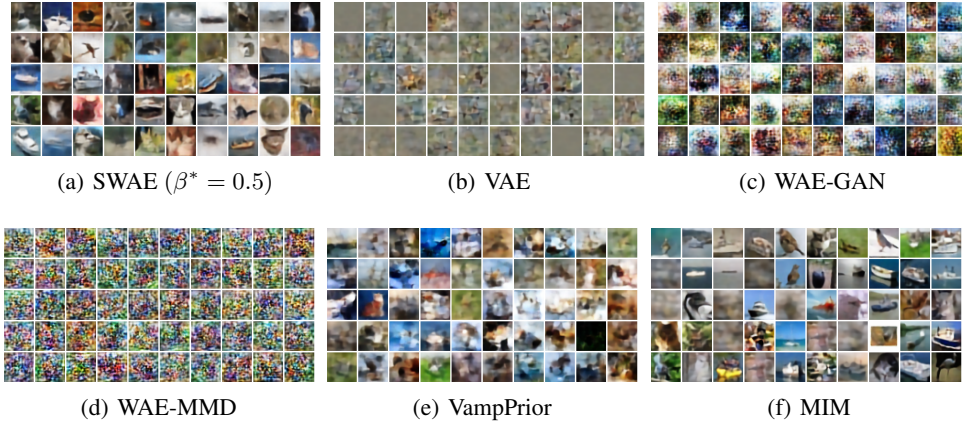
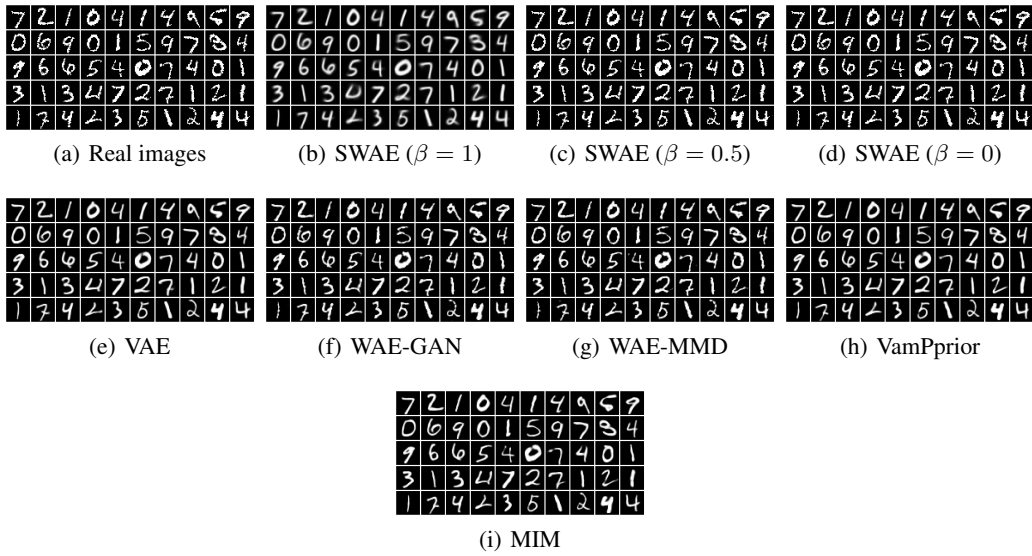
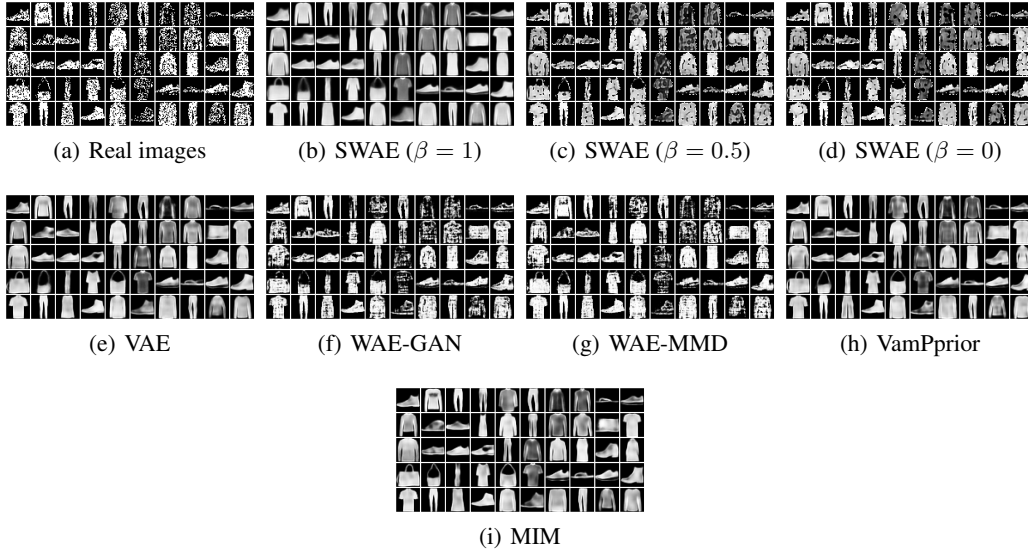
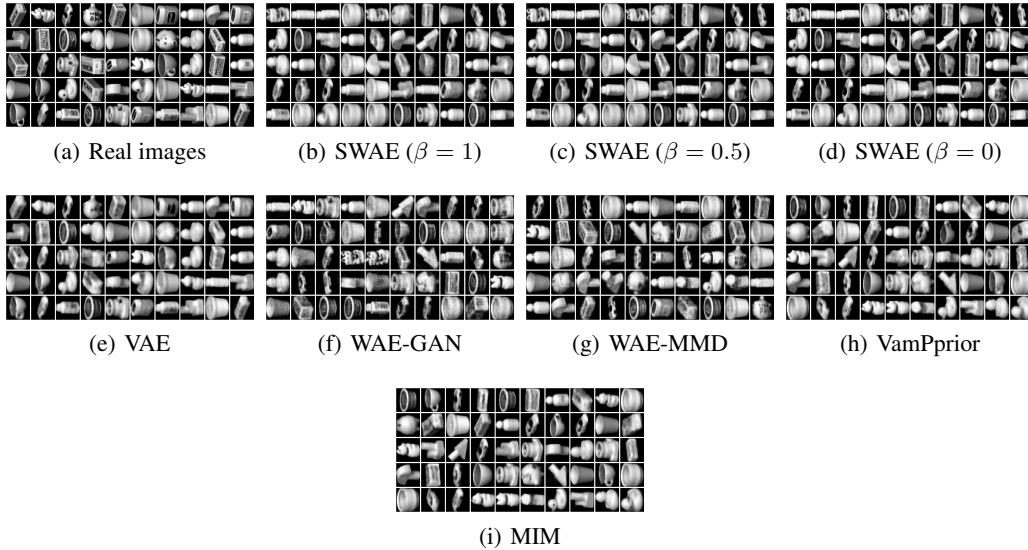


Figure 6: Generated new samples on Fashion-MNIST.  $\dim\mathbf{z} = 8$  for all methods.

Figure 7: Generated new samples on Coil20.  $\dim\mathbf{z} = 80$  for all methods.Figure 8: Generated new samples on CIFAR10-sub.  $\dim\mathbf{z} = 512$  for all methods.Figure 9: Reconstructed images on MNIST.  $\dim\mathbf{z} = 80$  for all methods. As expected, for SWAEs a smaller  $\beta$  leads to a higher quality of reconstruction.



Figure 10: Reconstructed images on Fashion-MNIST.  $\dim-z = 80$  for all methods.Figure 11: Reconstructed images on Coil20.  $\dim-z = 80$  for all methods. For SWAEs, the difference of the reconstruction error for different values of  $\beta$  is insignificant, and the reconstructed images look visually the same.

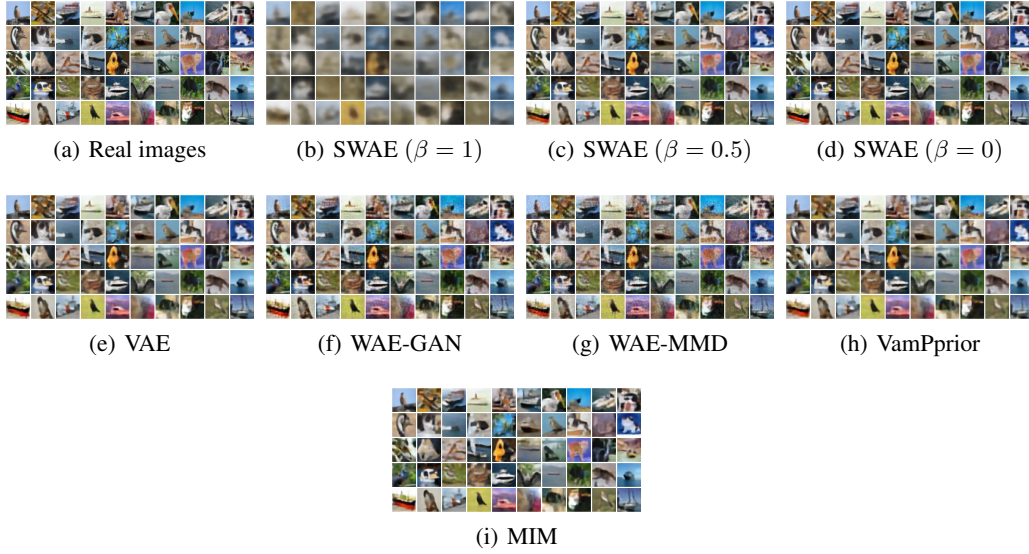


Figure 12: Reconstructed images on CIFAR10-sub.  $\dim-z = 512$  for all methods. Excluding the reconstruction loss in the objective, the reconstruction of SWAE ( $\beta = 1$ ) is blurry.

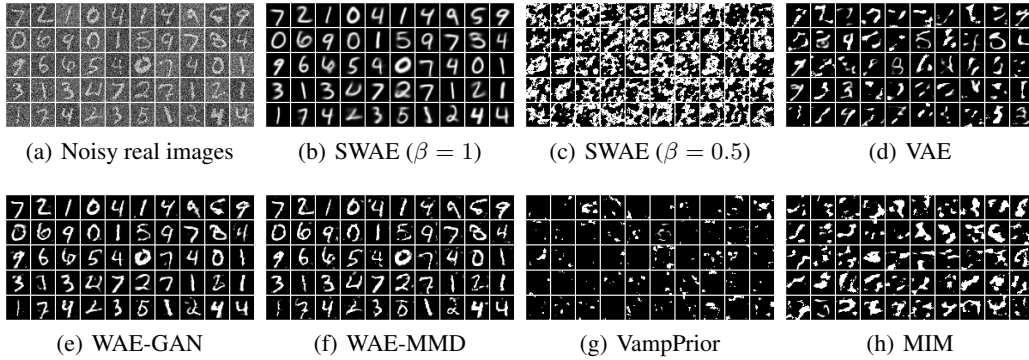


Figure 13: Denoising effect: reconstructed images on MNIST.  $\dim-z = 80$  for all methods. SWAE ( $\beta = 1$ ), WAE-GAN, and WAE-MMD can recover clean images. However, for WAE-GAN and WAE-MMD, we can still see some noisy dots around the digits.

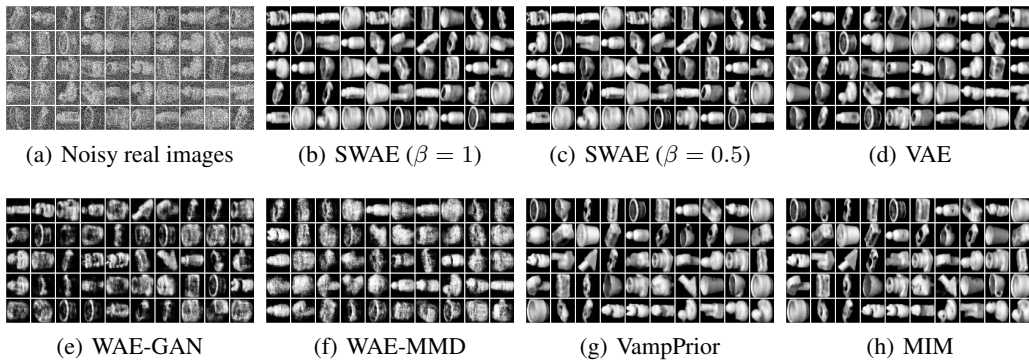


Figure 14: Denoising effect: reconstructed images on Coil20.  $\dim-z = 80$  for all methods. Except WAE-GAN and WAE-MMD, the other methods can produce clear images.