A Decoder Suffices for Query-Adaptive Variational Inference

Abstract

Deep generative models like variational autoencoders (VAEs) are widely used for density estimation and dimensionality reduction, but infer latent representations via amortized inference algorithms, which require that all data dimensions are observed. VAEs thus lack a key strength of probabilistic graphical models: the ability to infer posteriors for test queries with arbitrary structure. We demonstrate that many prior methods for imputation with VAEs are costly and ineffective, and achieve superior performance via query-adaptive variational inference (QAVI) algorithms based directly on the generative decoder. By analytically marginalizing arbitrary sets of missing features, and optimizing expressive posteriors including mixtures and density flows, our non-amortized QAVI algorithms achieve excellent performance while avoiding expensive model retraining. On standard image and tabular datasets, our approach substantially outperforms prior methods in the plausibility and diversity of imputations. We also show that QAVI effectively generalizes to recent hierarchical VAE models for high-dimensional images.

1 INTRODUCTION

Structured probabilistic models, like graphical models (Koller and Friedman, 2009), provide the foundation for a wide range of machine learning applications. Formalizing the data generation process facilitates the incorporation of domain knowledge and enables unsupervised learning from unlabeled data. Probabilistic models also enable diverse inference queries by users, in which arbitrary subsets of features may be predicted given novel missing data patterns. A number of inference algorithms (Pearl, 1988; Dechter, 2003; Koller and Friedman, 2009) have been developed for models with discrete or Gaussian latent variables, which efficiently compute marginals of query variables given heterogeneous observations. While exact posterior inference is intractable for models with rich dependencies, optimization-based variational methods (Wainwright and Jordan, 2008) often provide effective approximations.

Variational bounds were classically optimized via *coordinate ascent variational inference* (CAVI, Jordan et al. (1999)) algorithms that iteratively update posterior approximations for individual variables. CAVI updates are effective for many parametric models composed from conjugate priors, and can have efficient message-passing implementations (Ghahramani and Beal, 2001; Winn and Bishop, 2005). But, CAVI updates are based on integrals that often lack closed forms, requiring Monte Carlo approximations (Paisley et al., 2012; Kucukelbir et al., 2017) of uncertain quality.

Stochastic subsampling of data helps scale variational learning to big datasets (Hoffman et al., 2013), but iterative CAVI updates may still be slow for complex models. *Amortized* variational inference (Mnih and Gregor, 2014) seeks to boost training efficiency by determining variational posteriors via an inference (or recognition) network, which is shared (or amortized) across many similar inference tasks. *Variational autoencoders* (VAEs, Kingma and Welling (2014); Rezende et al. (2014)) are deep generative models that utilize amortized inference to jointly train a generative "decoder" and inference "encoder". Sophisticated generalizations to the encoder and decoder networks (Kingma and Welling, 2019; Sønderby et al., 2016; Vahdat and Kautz, 2020a; Child, 2021a) have produced hierarchical VAEs that realistically model complex image data via dozens of stochastic layers.

While amortized inference has enabled the learning of impressive deep generative models, it sacrifices the flexibility of CAVI to handle arbitrary inference queries. Because VAEs are typically trained from fully-observed data, the inference encoder implicitly assumes *complete* and *uncorrupted* observations of every data dimension (e.g., pixel) are available. Some prior work (see Sec. 2) has used simple heuristics for learning VAEs with missing data (Mattei and

Submitted to the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023). **To be used for reviewing only**.

Frellsen, 2019; Nazábal et al., 2020; Collier et al., 2020), such as filling missing features with zeros. However, we show that in addition to requiring expensive encoder retraining for peak performance, these approaches are inaccurate unless the data and test inference queries are very simple or known in advance (during model training).

Amortized inference networks also produce sub-optimal variational bounds, and this "amortization gap" may be significant, especially for complex datasets (Cremer et al., 2018; Krishnan et al., 2017). While quick-and-approximate inference may be sufficient to provide a noisy gradient signal in the midst of a long training process, it is problematic when applied to test queries, especially in domains like medicine where accurate uncertainty quantification is crucial.

To address these challenges, we propose *query-adaptive variational inference* (QAVI) methods that approximate the posterior of missing data with arbitrary patterns, given only a trained generative decoder and sparse observations. Our QAVI approach has the same inferential flexibility as classic CAVI algorithms, and critically does not require a database with many examples of the missing-data pattern of interest. But unlike classic CAVI algorithms, QAVI is applicable to any (differentiable) model with continuous latent variables, including deep generative models like hierarchical VAEs. While some prior work has improved VAE training by reducing amortization gaps (Kim et al., 2018; Marino et al., 2018), our application of non-amortized inference to missing-data queries for deep generative models is novel.

We begin in Sec. 2 by reviewing prior work on handling missing data with (hierarchical) VAEs. Sec. 3.1 then develops QAVI algorithms that optimize variational bounds for the missing feature values, rather than filling them via greedy heuristics. In Sec. 3.2, we develop an alternative QAVI algorithm that directly optimizes the posterior of the latent data encoding, without amortization. Doing this allows exact marginalization of missing feature values, and enables flexible posterior approximations including mixture models (Jaakkola and Jordon, 1999; Gershman et al., 2012) and normalizing flows (Rezende and Mohamed, 2015). Results in Sec. 4 then show substantial qualitative and quantitative improvements in capturing multimodal posterior uncertainty for VAE models of images and tabular data. Here, we also show that our approach generalizes to state-of-the-art hierarchical VAE models of images (Child, 2021a).

2 BACKGROUND AND RELATED WORK

2.1 THE VARIATIONAL AUTOENCODER (VAE)

VAEs model the distribution of typically high-dimensional observed data x using continuous, lower-dimensional latent variables z, via the following generative model:

$$z \sim p(z), \qquad x \sim p_\theta(x \mid z). \tag{1}$$

For standard VAEs, the latent code $z \in \mathbb{R}^d$ has a factorized Gaussian prior $p(z)$. Given z, data is generated via a *decoder* (deep) neural network with weights θ. The decoder maps z to a likelihood $p_\theta(x|z)$ which may be a factorized Gaussian, or a distribution for constrained data like the continuous Bernoulli (Loaiza-Ganem and Cunningham, 2019).

The VAE log-likelihood $\log p_\theta(x) = \log \int p_\theta(x|z)p(z)\,dz$ is intractable. Learning thus typically employs *amortized* VI, where an *encoder* with parameters ϕ approximates the posterior over latent codes $q_\phi(z|x) \approx p_\theta(z|x)$. We jointly learn θ, ϕ by maximizing the *evidence lower-bound* (ELBO):

$$\mathcal{L}(\theta, \phi; x) = E_{q_\phi(z|x)}[\log p_\theta(x|z)] - \mathrm{KL}(q_\phi(z|x)||p(z)) \tag{2}$$

where KL is the Kullback–Leibler divergence, calculated analytically when $q_\phi(z|x)$ and $p(z)$ are both Gaussian. The ELBO provides a lower bound on the evidence $\log p_\theta(x)$ that is tight when the variational posterior $q_\phi(z|x)$ is exact. This expectation can be approximated via Monte Carlo samples from $q_\phi(z|x)$. Gradients with respect to θ, ϕ can then be estimated by the reparameterization "trick" of sampling from $q_\phi(z|x)$ via linear transforms of standard normal variables (Kingma and Welling, 2014; Rezende et al., 2014).

2.2 HIERARCHICAL VAES

Hierarchical VAEs (HVAEs, Sønderby et al. (2016); Klushyn et al. (2019)) extend the VAE by partitioning the latent code into L disjoint groups $z = (z_1, z_2, ...z_L)$, increasing the expressiveness of the generative model for complex data like images (Vahdat and Kautz, 2020b; Child, 2021b). HVAEs generate these stochastic codes sequentially as $p_\theta(x|z) = p_\theta(z_1)(\prod_{\ell=2}^{L} p_\theta(z_\ell|z_{<\ell}))p_\theta(x|z_L)$, with a similar encoder: $q_\phi(z|x) = q_\phi(z_1|x)\prod_{\ell=2}^{L} q_\phi(z_\ell|z_{<\ell}, x)$. Each conditional in the decoder $p_\theta(z_\ell|z_{<\ell})$, and the encoder $q_\phi(z_\ell|z_{<\ell})$, is typically Gaussian with mean and variance determined by (non-linear) neural networks.

The HVAE ELBO (Vahdat and Kautz, 2020b) equals

$$\mathcal{L}_H(\theta, \phi; x) = E_{q_\phi(z|x)}[\log p_\theta(x|z)] - \mathrm{KL}(q_\phi(z_1|x)||p_\theta(z_1)) - \sum_{\ell=2}^{L} E_{q_\phi(z_{<\ell}|x)}[\mathrm{KL}(q_\phi(z_\ell|z_{<\ell}, x)||p_\theta(z_\ell|z_{<\ell}))], \tag{3}$$

where $q_\phi(z_{<\ell}|x) = \prod_{i=1}^{\ell-1} q(z_i|z_{<i}, x)$ is the approximate posterior up to latent group $(\ell - 1)$. Reparameterization is then used to provide Monte Carlo gradient estimates.

We can rewrite the conditional prior and approximate posterior for layer ℓ to make the set of relevant networks explicit:

$$p_\theta(z_\ell|z_{<\ell}) = \mathcal{N}(z_\ell|\ \mu_{\theta_\ell}(z_{<\ell}),\ \sigma_{\theta_\ell}(z_{<\ell})), \tag{4}$$
$$q_\phi(z_\ell|z_{<\ell}, x) = \mathcal{N}(z_\ell|\ \mu_{\phi_\ell}(f_{\phi_\ell}(x), g_{\phi_\ell}(z_{<\ell})),\ \sigma_{\phi_\ell}(...)).$$

Here, f_{ϕ_ℓ} and g_{ϕ_ℓ} are networks that extract feature representations of the observation x and the previous layers $z_{<\ell}$, respectively. These features determine the mean and scale of the conditional Gaussian posterior via $\mu_{\phi_\ell}, \sigma_{\phi_\ell}$. Networks $\mu_{\theta_\ell}, \sigma_{\theta_\ell}$ similarly generate the prior parameters for layer ℓ.

2.3 INFERENCE QUERIES FOR MISSING DATA

After training, synthetic data may be easily generated from $p_\theta(x)$ by sampling z from the learned VAE or HVAE prior, and $x \sim p_\theta(x|z)$. However, the learned encoder does *not* provide a direct mechanism for conditional queries about missing features, given partial or corrupted test data.

Let $x = (x_O, x_M)$ be a test data point with observed features x_O and missing features x_M. We are particularly interested in scenarios where the specific feature dimensions that are missing and observed will vary across test instances. Given a trained (hierarchical) VAE, for which x_O and x_M are conditionally independent given z, missing data is optimally predicted via the conditional distribution:

$$p_\theta(x_M \mid x_O) = \int p_\theta(x_M \mid z) p_\theta(z \mid x_O)\, dz. \quad (5)$$

Exactly evaluating this distribution is challenging due to the non-linearity of the decoder, and the intractability of the posterior of latent codes z. In recent years a number of approximations have been proposed to address this problem.

Heuristic Preprocessing. Imputation heuristics, such as replacing missing features with statistical summaries like their mean or mode, are common in machine learning and statistics. Some recent work on training VAEs in the presence of missing data (Mattei and Frellsen (2019); Nazábal et al. (2020); Collier et al. (2020)) propose a *Fill-Zeros* heuristic, simply replacing missing features with zeros as input to the VAE encoder. While Fill-Zeros may be effective for learning models of MNIST digits (where zeros are common) when pixels are missing uniformly at random, our experiments show that in even slightly more complex scenarios, its performance is very poor.

Monte Carlo methods. Rezende et al. (2014) first proposed a simple scheme to approximately sample from $p_\theta(x_M \mid x_O)$ by starting with a random imputation, which is then stochastically encoded and decoded (or autoencoded) several times. Because the encoder only approximates the true posterior distribution $p_\theta(z \mid x)$, this *pseudo-Gibbs* sampler will not sample from the true posterior of missing features, and encoder inaccuracies may cause its equilibrium distribution to be far from $p_\theta(x_M \mid x_O)$. This approach was improved by Mattei and Frellsen (2018), who proposed a Metropolis-Hastings correction to each step of the pseudo-Gibbs sampler, inducing a *Metropolis-in-Gibbs* sampler that asymptotically samples from $p_\theta(x_M \mid x_O)$. While Metropolis-in-Gibbs converges to the true posterior, it does so at a rate that may be impractically slow.

Amortized Inference for Imputation. To improve on heuristic preprocessing methods, recent works have introduced the concept of learning a new "partial" encoder approximating $p_\theta(z \mid x_O)$ with the help of a pre-trained VAE. Collier et al. (2020) aims to do this by concatenating zero-filled data with a binary mask indicating missing features, and optimizing the standard VAE ELBO of Eq. (2). While this approach is specifically developed to *train* VAEs with missing data, it is trivially applicable to test-time missing data imputation by retraining the encoder with the masked test data, generating a *Re-tuned Encoder* before evaluation. Alternatively, *Posterior Matching* (Strauss and Oliva, 2022) artificially masks the fully-observed training data x as x_O, and tunes the masked encoder $q_\psi(z \mid x_O)$ to maximize the objective: $E_{z\sim q_\phi(z|x)}[\log q_\psi(z|x_O)]$. Finally, Harvey et al. (2022) propose *Inference in a Pretrained Artifact* (*IPA*) for imputation with HVAEs. IPA optimizes a similar loss to Posterior Matching, but also fine-tunes the generative model to produce an approximate conditional model $q(x_M|x_O)$.

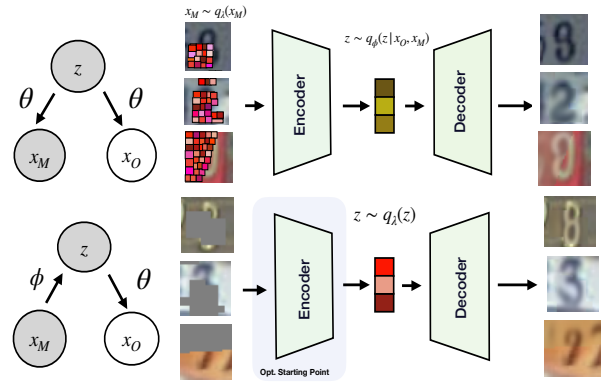

Figure 1: Overview of QAVI. **Top** queries with missing features $q_\lambda(x_M)$ directly and **bottom** queries via latent space, $q_\lambda(z)$. QAVI can be integrated to any existing VAE since θ and ϕ are fixed. For both top and bottom, **left** shows the dependencies among the variables and **right** shows QAVI's implementation using components from a pre-trained VAE and additional variational parameters λ being optimized (shown in red). After fitting $q_\lambda(x_M)$ (in top figure) using the VAE, we can draw multiple samples $x_M \sim q_\lambda(x_M)$ and pass $x = \{x_O, x_M\}$ through the VAE to compute each query $p_\theta(x_M|x_O)$. Similarly, after fitting $q_\lambda(z)$ (in bottom figure) we can decode samples from it, re-using only the decoder and throwing the decoder away.

These amortized approaches to missing data imputation with VAEs have several drawbacks compared to our QAVI method. *1)* They incur a substantial initial overhead for training the partial encoder. *2)* Partial encoder training requires access to a relatively large set of partially-observed examples, and/or continued access to the training set along with a known missing-feature distribution. *3)* They are sensitive to shifts in the distribution of queries (missing-feature patterns) between training and evaluation time. If evaluated on previously unseen missing-feature patterns, performance may suffer substantially (see Fig. 4). *4)* As we will demonstrate, even in the ideal setting, performance can be sub-optimal.

3 QUERY ADAPTIVE VI

In this section, we introduce *QAVI*, which utilizes the pre-trained VAE as prior knowledge, defines variational free parameters for each query, and refines them. We give an

overview in Fig. 1 and show that it is adaptive across different queries without the need for additional inference networks. QAVI lets us define a variational posterior over the unobserved variables $q_\lambda(z, x_M)$ with parameters λ. Fig. 1 shows the dependencies between these two unobserved variables, resulting in different interpretations of the variational parameters, individually elaborated below.

3.1 VARIATIONAL POSTERIORS FOR MISSING FEATURES

Suppose first that the variational posterior factorizes as

$$q_\lambda(x_M, z) = q_\lambda(x_M) q_\lambda(z|x_M, x_O).$$

Doing this directly expresses the uncertainty in the posterior distribution of missing pixels (see Figure 1) and their impact on the latent code. We substitute $q_\lambda(z|x_M, x_O)$ with the amortized, pre-trained encoder $q_\phi(z|x)$, assuming that the encoder optimally represents the relationship between x and z. In our experiments, we found that fitting $q_\lambda(x_M)$ using the ELBO (see supplements for derivation) resulted in low variance posteriors and hence, we employ hyper-parameters β and γ, as in (Higgins et al., 2017), to capture the uncertainty. The variational lower bound $\mathcal{L}_M$ becomes :

$$\mathcal{L}_M(\lambda; x) = E_{q_\lambda(x_M)}\big[E_{q_\phi(z|x_O,x_M)}[\log p_\theta(x_O, x_M|z)] - \beta * \mathrm{KL}(q_\phi(z|x_M, x_O)||p(z))\big] + \gamma * H(q_\lambda(x_M)) \quad (6)$$

where H is the entropy. We approximate this lower bound with S Monte-Carlo samples from $q_\lambda(x_M)$, as in conventional VAE training:

$$\mathcal{L}_M(\lambda; x) \approx \frac{1}{S}\sum_{s=1}^{S}\big[\log p_\theta(x_O, x_M^{(s)}|z^{(s)}) - \mathrm{KL}(q_\lambda(z|x_O, x_M^{(s)})||p(z))\big] + H(q_\lambda(x_M)) \quad (7)$$

where $x_M^{(s)} \sim q_\lambda(x_M)$, and $z^{(s)} \sim q(z|x_O, x_M^{(s)})$. Note that the KL and H terms are calculated in closed form. Automatic differentiation may then be used to compute gradients with respect to the posterior parameters λ, coupling the impact of missing features on the observed data indirectly.

While this approach correctly captures uncertainty in the posterior distribution of missing pixels, its optimization requires repeated computation of the encoder and decoder. It also suffers from any inaccuracies in the trained encoder. Perhaps surprisingly, we will show that in this context, a fully non-amortized inference method can actually have greater computational efficiency and accuracy.

3.2 NON-AMORTIZED VARIATIONAL INFERENCE

We can construct an alternative variational posterior via the following factorization (Figure 1):

$$q_\lambda(x_M, z|x_O) = q_\lambda(z) q_\lambda(x_M|z),$$

By defining a variational posterior on z, we no longer make use of the encoder (except for initialization) and re-use the pre-trained decoder by specifying $q_\lambda(x_M|z) = p_\theta(x_M|z)$. Our non-amortized variational objective $\mathcal{L}_N$ becomes (derivation in supplements):

$$\mathcal{L}_N(\lambda; x) = E_{q_\lambda(z)}[\log p_\theta(x_O|z)] - \beta * \mathrm{KL}(q_\lambda(z)||p(z)). \quad (8)$$

The above non-amortized ELBO only explicitly involves sampling from the latent code, marginalizing the missing data. This approach finds the optimal latent space parameters that best explain the partially observed data x_O while also keeping them close to the prior with a factor of β (> 1). In the simplest case, we define $q_\lambda(z)$ to be diagonal Gaussian. Eq. 8 can be estimated with S monte carlo samples from $q_\lambda(z)$ and λ (mean and variance) optimized with stochastic gradient descent.

Missing queries can induce multi-modal uncertainties which are not captured by the single mode Gaussian distribution. To address this, we extend QAVI to more expressive variational distributions that may better approximate the true posterior.

Flow Posteriors. The flow-based variational posterior aims to construct a complex distribution by transforming a simple Gaussian through a series of invertible mappings. We let $z_T = \prod_{t=1}^{T} \mathcal{T}_t(z_{t-1}, \lambda_t)$, where z_0 is the variable corresponding to the base distribution with parameters λ_0. λ_t is the set of parameters specified by each flow layer $\mathcal{T}_t$ and T is the total number of such transformations. The full set of variational parameters is $\lambda = \{\lambda_0, \lambda_1, ..., \lambda_T\}$. The idea of improving AVI in VAEs with the help of such *Normalizing Flows* was first introduced by Rezende and Mohamed (2015). We instead, employ autoregressive transformations to capture high-dimensional dependencies in the latent space, named Inverse Autoregressive Flows (IAF) (Kingma et al., 2016). We estimate both terms in $\mathcal{L}_N(\lambda; x)$ (Eq. (8)) by Monte Carlo samples; we draw multiple samples from $q_\lambda(z_T)$, first by sampling from the Gaussian and transforming them via the flow layers. λ is finally optimized by performing stochastic backpropagation.

Gaussian Mixture Posteriors. Parameterizing the latent space distribution $q_\lambda(z)$ as a *mixture of (diagonal) Gaussians* enables us to explicitly model different hypothesis in the latent space. In this case, $q_\lambda(z) = \sum_{i=1}^{N} w^{[i]} \mathcal{N}(z|\mu^{[i]}, \sigma^{[o]})$, where $\mu^{[i]}$, $\sigma^{[i]}$ are the individual mixture component means and standard deviation respectively and $w^{[i]}$ are the mixture weights (with $\sum_{i=1}^{N} w^{[i]} = 1$). Consequently, we define $\lambda = \{\mu^{[1]}, ..., \mu^{[N]}\} \cup \{w^{[1]}, ..., w^{[N]}\} \cup \{\sigma^{[1]}, ..., \sigma^{[N]}\}$. Optimizing λ is not straight-forward as the "reparameterization trick" does not extend to mixture models, since it is difficult to reparameterize the discrete distribution over mixture weights. Alternatively, we use Implicit Reparameterization Gradients (Graves (2016), Figurnov et al. (2018)) to optimize the

means and variances of individual components and employ Importance Sampling Gradient estimator (Ścibior and Wood, 2021) for computing gradients w.r.t. the mixture weights and perform λ updates via Eq. 8 (refer to supplements for a detailed description on the gradients).

Hierarchical VAE Posteriors. For the hierarchical VAE models (Sønderby et al., 2016) introduced in section 1, neither Gaussian nor mixture of Gaussian distributions are flexible enough to capture the nonlinear dependencies between latent variables at different levels of the hierarchy. We therefore propose a new more expressive (non-amortized) variational family for hierarchical VAEs that removes the dependency on the observation (x), while retaining the expressive, nonlinear dependency structure of the original model. Our proposed distribution factorizes as:

$$q_\lambda(z) = q_\lambda(z_1) \prod_{\ell=2}^{L} q_\lambda(z_\ell | z_{<\ell}). \tag{9}$$

As in the original model, we let $q_\lambda(z_\ell|z_{<\ell})$ be conditionally Gaussian for all l. More complex conditional distributions (Flow, Mixtures, etc.) could also be used, but this conditionally Gaussian structure alone allows for expressive, multi-modal posterior approximations.

We introduce a simple, generic strategy for constructing a family of non-amortized distributions of this form given a pre-trained hierarchical VAE. Our approach applies to many recent hierarchical VAE architectures, including the "very-deep" architecture of Child (2021b) that we use in our experiments. To specify the non-amortized QAVI distribution we begin with the amortized posterior defined in equation 4. Holding g_{ϕ_ℓ}, μ_{ϕ_ℓ}, σ_{ϕ_ℓ} fixed, we replace the features extracted from the observation with a new tuneable parameter λ_ℓ. A further set of weighting parameters, $\gamma_\ell \in [0, 1]$, weight the output parameters against those of the prior. Thus our hierarchical QAVI posterior for layer ℓ becomes:

$$\begin{aligned} \mu_\ell &= \gamma_\ell \cdot \mu_{\phi_\ell}(\lambda_\ell, g_{\phi_\ell}(z_{<\ell})) + (1-\gamma_\ell) \cdot \mu_{\theta_\ell}(z_{<\ell}) \\ \sigma_\ell &= \gamma_\ell \cdot \sigma_{\phi_\ell}(\lambda_\ell, g_{\phi_\ell}(z_{<\ell})) + (1-\gamma_\ell) \cdot \sigma_{\theta_\ell}(z_{<\ell}) \\ q_\lambda(z_\ell|z_{<\ell}) &= \mathcal{N}(z_\ell|\ \mu_\ell,\ \sigma_\ell) \end{aligned} \tag{10}$$

Re-using the pre-trained networks g_{ϕ_ℓ}, μ_{ϕ_ℓ} and σ_{ϕ_ℓ} allows the full amortized encoder to be used for initialization of λ_ℓ by simply setting $\lambda_\ell \leftarrow f_{\phi_\ell}(x_O, \tilde{x}_M)$. $\tilde{x}_M$ may be any initialization for the missing features, even Gaussian noise.

Our approach of interpolating posterior (μ_{ϕ_ℓ}, σ_{ϕ_ℓ}) and prior (μ_{θ_ℓ}, σ_{θ_ℓ}) network outputs is vital when reusing μ_{ϕ_ℓ} and σ_{ϕ_ℓ} from the original inference model. In the original, fully observed, training phase posterior variances may become extremely tight for the highly overparameterized HVAE model. Conversely, with missing data, latent variables corresponding to unobserved features should have distributions close to the prior. Expressing the missing data posterior as a weighted combination of prior and posterior network outputs allows our variational family to easily produce appropriate posteriors for latent variables corresponding to both obsevered and unobserved portions of the image, without needing to re-tune μ_{ϕ_ℓ} and σ_{ϕ_ℓ} .

Figure 2: Comparison of sampled imputations for the FFHQ-256 dataset from QAVI, Posterior Match and IPA. Samples were taken with temperature 0.85 (see Child (2021b).

Hierarchical VAE Warmup For hierarchical VAEs with a multi-scale architecture, we find that a warmup phase of optimization with a modified objective greatly accelerate posterior fitting. Our modified objective is:

$$\begin{aligned} &\mathcal{L}_H(\theta, \phi; x_O) = \\ &\quad E_{q_\lambda(z)}[\log p_\theta(x_O|z)] - \frac{1}{d_1}\mathrm{KL}(q_\lambda(z_1)||p_\theta(z_1)) \\ &\quad - \sum_{\ell=2}^{L} \frac{1}{d_\ell} E_{q_\lambda(z_{<\ell})}[\mathrm{KL}(q_\lambda(z_\ell|z_{<\ell})||p_\theta(z_\ell|z_{<\ell}))], \end{aligned} \tag{11}$$

where d_ℓ is proportional to the size of the latent space at layer ℓ. Intuitively, when inpainting large, continuous regions of an image, high-level representations should be determined first, with details filled later. However, in the unmodified ELBO higher resolution latent layers contribute substantially more to the loss, leading to slow convergence.

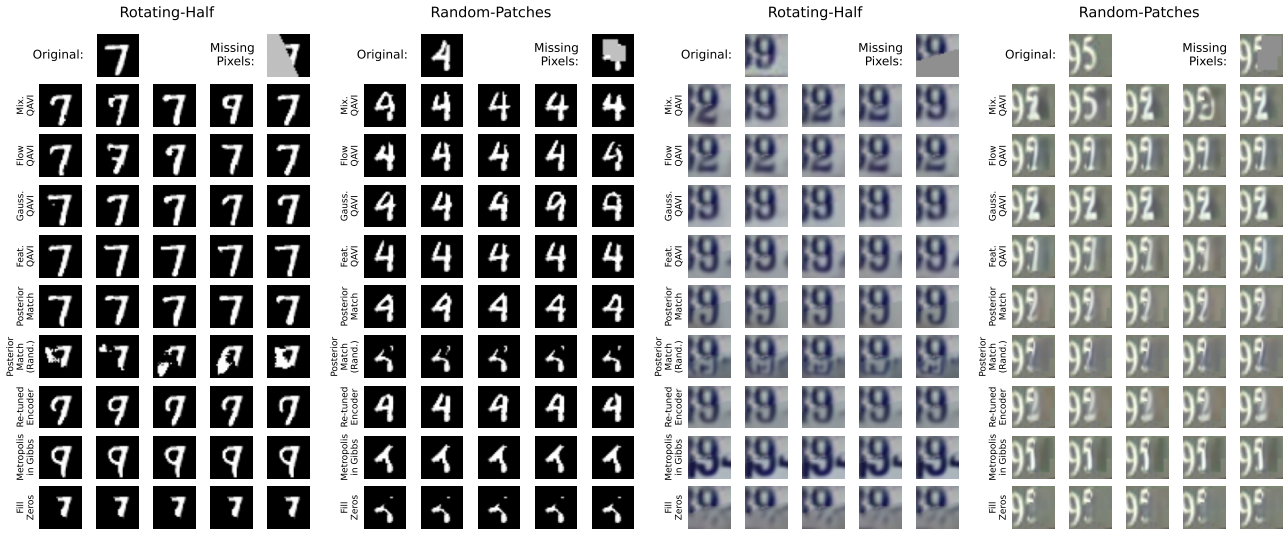

Figure 3: Digit completion results on real-valued MNIST (**left**) and SVHN (**right**) images for randomized queries of the Rotating-Half and Random-Patches of missing pixels. The plot shows 5 samples from each inference procedure. Fill zeros fail to infer any valid values for missing features. Monte Carlo Simulations and Amortized Inference methods propose one valid digit completion, with AVI showing samples closer to the true digit. We see that the performance among the different Posterior Match settings vary quite widely depending on the missing distribution it is trained on. We see that QAVI adapts to each query and proposes not only valid imputations closer to the true digit, but also with expressive posteriors defined by Flow and Mix. QAVI, it can capture the uncertainty in the digits.

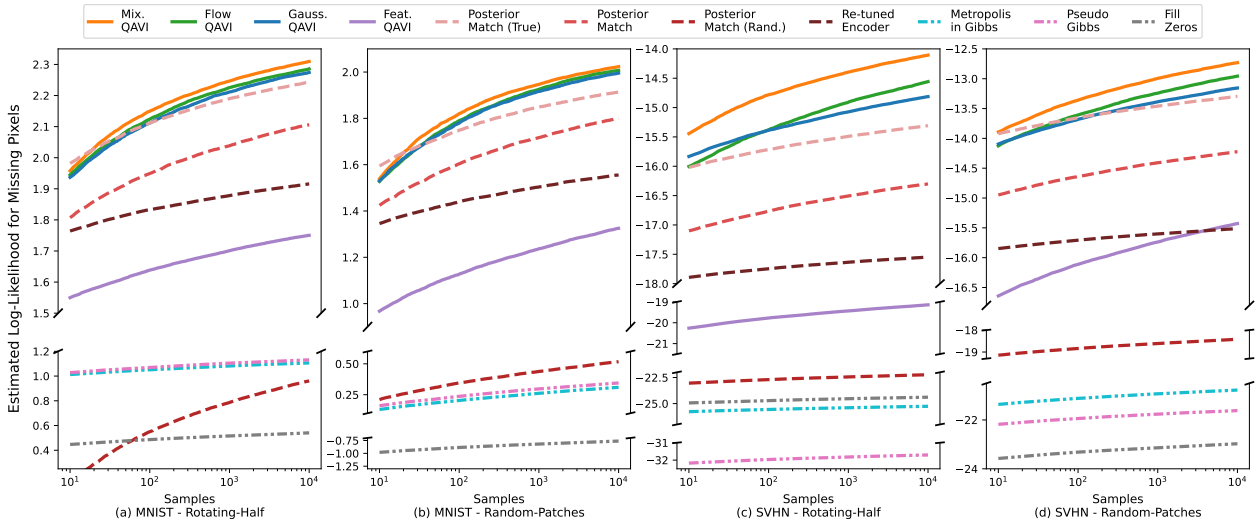

Figure 4: The plots show the estimated log-likelihood for missing pixels of various inference methods for 1000 randomly sampled MNIST and SVHN test images as a function of the number of samples used.

4 EXPERIMENTS & RESULTS

4.1 EXPERIMENTAL SETUP

We evaluate QAVI with 6 tabular datasets from the UCI Machine Learning Database (Dua and Graff, 2017). For tabular data, we follow the experimental setup of Mattei and Frellsen (2019) to train our VAEs, but with a Gaussian variational family instead of Student's t. We also consider 3 image datasets: real-valued MNIST (LeCun and Cortes (2010)), Street View House Number dataset (SVHN) (Netzer et al., 2011), FFHQ-256 (Karras et al., 2019) and 6 tabular datasets from UCI Machine Learning Database (Dua and Graff, 2017). We consider a single-stochastic-layered VAE for MNIST and SVHN, with a WideResNet architecture (Zagoruyko and Komodakis, 2016) that is known to work well with images. We train them with fully-observed images from the training set ($70,000$ MNIST images and $73,257$ SVHN images) and maximize the variational lower bound of Eq. (2). For FFHQ-256, we adopt the "very-deep"

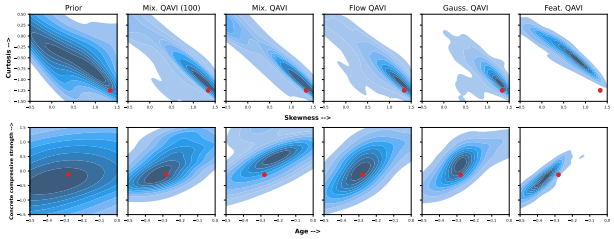

Figure 5: Kernel Density estimate plots for the prior and 5 QAVI variants, fit using 1000 samples from the inferred missing posterior, when one test sample chosen from two UCI datasets, Banknote (**top**) and Concrete (**bottom**), contain exactly two missing features (labeled as the respective axes). We observe that the true data point, marked in red, is often enclosed by the different posteriors, but sometimes missed by Feat. QAVI. We also visualize the posterior using a highly expressive mixture (Mix. QAVI (100)), with 100 Gaussian components.

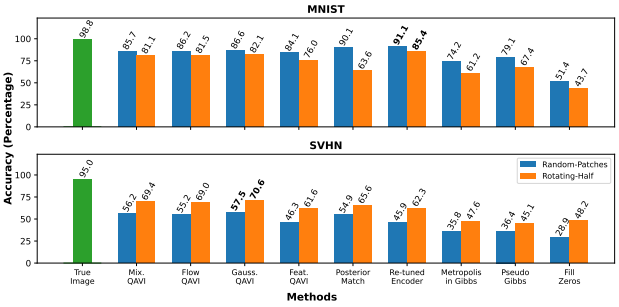

Figure 6: % Classification accuracies using 100 samples from the inferred missing posterior across inference methods for randomized missing queries in MNIST (left) and SVHN (right). We use a trained discriminative model (using WRN-28-2 architecture (Zagoruyko and Komodakis (2016))) to predict class labels.

hierarchical architecture of Child (2021b), but for efficient comparison, we re-trained a smaller variant of the original FFHQ-256 model (5.9M vs. 115M parameters).

For our experiments with MNIST and SVHN, we consider the following missing-at-random (MAR) patterns: 1) two randomly placed patches (each of size 10x10 for MNIST, 15x15 for SVHN), 2) randomly rotated mask of half of the image. We consider a more challenging randomized masking scheme (Zhao et al., 2021) for FFHQ-256. For tabular datasets, we corrupt the test set by removing half of the features in each row uniformly at random.

Our experimental setup is consistent with most existing work on imputation with VAEs. We do not consider missing-not-at-random (MNAR) data (Ipsen et al., 2020), but we believe our QAVI methods may provide the foundation for future advances in MNAR inference.

Baselines. We compare QAVI with several methods from recent literature: *i)* The "Fill Zeros" heuristic (Nazábal et al., 2020); *ii)* Monte Carlo methods: "pseudo-Gibbs" (Rezende et al., 2014) and "Metropolis-in-Gibbs" (Mattei and Frellsen, 2018); *iii)* Amortized inference methods: "Re-tuned Encoder" (Collier et al., 2020) and "Posterior Match[ing]" (Strauss and Oliva, 2022).We consider 3 variants of posterior matching to evaluate the importance of knowing the missing-feature pattern prior to training. "Posterior Match (True)" indicates that the posterior matching encoder is trained with exactly the same query distribution used for test evaluation. "Posterior Match (Rand.)" assumes that only the fraction of missing features is known; features are removed at random with probability equal to this fraction. For the generic "Posterior Match" we use the image masking distribution of Zhao et al. (2021), which assumes contiguous masked regions without specific knowledge of queries to be evaluated.

For HVAEs we compare QAVI with state-of-the-art approaches to inpainting with HVAEs: "Posterior Match[ing]" and "IPA" (Harvey et al., 2022). Both "Posterior Match" and "IPA" were trained using the same random masking distribution (Zhao et al., 2021) used in evaluation.

Hyperparameters. For QAVI optimization we use 100 Monte Carlo samples to estimate our variational objective.We further optimize the variational parameters λ using the Adam optimizer for 300 steps. Across experiments, we refer to our posterior on missing features as "Feat. QAVI" and our Gaussian/Flow/Mixture variational posterior on the latent space z as "Gaus./Flow/Mix. QAVI" respectively. For HVAE QAVI we use 28 Monte Carlo samples to estimate the objective and optimize for 1000 total steps (incl. 500 warmup steps). More architecture and training details can be found in the supplementary material.

Metrics. For a quantitative analysis, we estimate the marginal log-likelihood of the missing features x_M using the importance sampling estimator of Burda et al. (2015):

$$\log p_\theta(x_M) \geq \mathbb{E}_{z^{(1)}\ldots z^{(S)}\sim q(z)}\left[\log \frac{1}{S}\sum_{s=1}^{S}\frac{p_\theta(x_M, z^{(s)})}{q(z^{(s)})}\right],$$
$$\approx \log \frac{1}{S}\sum_{s=1}^{S}\frac{p_\theta(x_M, z^{(s)})}{q(z^{(s)})}; z^{(s)} \sim q(z). \quad (12)$$

While the true log-likelihood of missing features will be consistent across methods as the generative model $p_\theta(x)$ is fixed, better posterior approximations lead to tighter lower bounds for a fixed number of samples. Fig. 4 shows the estimated log-likelihoods versus the number of samples.

As the importance-weighted likelihood estimator becomes expensive and unreliable to evaluate for high-dimensional models like HVAEs, we use perceptual metrics to evaluate inpainting results for HVAEs on the FFHQ dataset. We report three metrics on a test set of 1000 images: FID (Heusel et al., 2017), P-IDS and U-IDS (Zhao et al., 2021). We modify the P-IDS and U-IDS metrics slightly to reduce sensitivity to the test set size; see supplement for details.

Table 1: Test missing data log-likelihoods (LL) and normalized root mean-square error (NRMSE) for 6 tabular datasets from the UCI repository, estimated using 10,000 samples. NRMSE per test row is the minimum mean-square error calculated across these 10,000 samples. Lower is better for MSE, and higher is better for LL. We highlight the best performing inference methods in bold, and observe that QAVI outperforms across datasets.

	Breast Cancer		Red wine		White wine		Banknote		Concrete		Yeast	
	LL	NRMSE	LL	NRMSE	LL	NRMSE	LL	NRMSE	LL	NRMSE	LL	NRMSE
Mix. QAVI	-9.16	0.39	**-6.35**	**0.26**	-7.03	0.35	-2.25	**0.1**	-2.70	**0.17**	2.42	**0.46**
Flow QAVI	-9.14	0.39	-6.47	**0.26**	-7.03	0.35	**-2.24**	**0.1**	**-2.65**	**0.17**	2.55	**0.46**
Gaus. QAVI	-9.21	0.39	-6.56	0.3	-6.94	**0.3**	-2.35	**0.1**	-2.82	**0.17**	**2.82**	**0.46**
Feat. QAVI	-16.32	0.37	-13.2	0.28	-8.8	0.38	-8.27	0.15	-15.14	0.23	-7.81	0.47
Posterior Match	-15.14	**0.33**	-14.83	0.42	-9.19	0.39	-4.42	0.14	-15.05	0.4	-430.36	0.47
Re-tuned Encoder	-12.82	0.4	-12.96	0.4	-9.59	0.39	-4.26	0.14	-11.48	0.35	-33.62	0.5
Metropolis-in-Gibbs	-23.94	0.44	-74.22	0.68	-19.83	0.58	-49.51	0.68	-265.62	0.62	-1.86	0.54
pseudo-Gibbs	-12.61	**0.33**	-33.53	0.36	-10.68	0.36	-44.16	0.4	-249.06	0.36	-5.11	0.48
Fill Zeros	-32.56	0.4	-21.97	0.36	-10.09	0.37	-17.46	0.32	-23.33	0.32	-11.27	0.48

Table 2: Quantitative comparison of perceptual inpainting quality on the FFHQ-256 dataset. We compare QAVI against two state-of-the-art adaptations of HVAEs to inpainting, using the same base "very-deep" HVAE architecture.

Method	FID ↓	P-IDS* ↑	U-IDS* ↑
QAVI	**21.21**	**6.20**	**24.98**
Posterior Match	23.68	3.36	21.55
IPA	26.41	2.31	18.19

4.2 RESULTS

QAVI improves imputation quality. Fig. 4 and Table 2 compare the log-likelihood of missing features across MNIST, SVHN, and tabular datasets. We see that heuristic imputation and Monte Carlo methods perform poorly. Retuned Encoder and Posterior Match show relatively higher likelihoods, but do not match QAVI across any dataset or missingness pattern. Feat. QAVI is competitive for tabular data, but for high-dimensional images it is susceptible to local optima. Fig. 6 similarly shows that QAVI provides reliably strong performance for downstream tasks.

QAVI can capture multi-modal posterior uncertainty. We show imputations for 4 test examples in Figure 3 to highlight differences in inference methods and to understand the uncertainty in the posterior over the missing features. We see that Gaussian and Flow QAVI defined on the latent space are capable of capturing uncertainty in the missing features. With a mixture of Gaussians variational family, QAVI produces multiple visually-plausible imputations. The classification performance in Fig. 6 shows high relative classification accuracy for expressive posteriors despite the increased variance in samples.

QAVI benefits extend to HVAEs. We find that QAVI integrates well with VAE models. The results in Table 2 show that our QAVI approach for HVAEs produces imputations with higher perceptual scores than prior methods for leveraging HVAEs for inpainting. Figure 2 shows that samples produced by QAVI are qualitatively more visually plausible. Further results in the supplement show the diversity of plausible samples.

Amortized imputation is sensitive to training queries. Fig. 4 compares the performance for Posterior Matching when trained with different masking distributions. We see that with the absence of any prior knowledge of the true masking distribution at train time, the performance of Posterior Matching can be as poor as simple heuristics like fill zeros. Even in the unrealistic case where the exact distribution of missingness is known at train time, posterior matching does not outperform QAVI. This sensitivity to the choice of missingness for training is significant: adaptation to new patterns of missingness require full retraining. In sensitive domains, such as medicine, access to the original training set may be restricted or even unavailable. QAVI is indifferent to the structure of queries, requires no retraining, and *still* outperforms the "best case" amortized imputation.

5 CONCLUSION

We have presented a simple and a general framework that has been unexplored in prior work applying VAEs to the imputation of missing data. Previous state-of-the-art approaches make use of a restrictive inference network as an imputation strategy. We instead: take an existing VAE decoder model for the joint distribution $p(x)$, allocate variational parameters on the latent variable z for every missing data point, and train the parameters stochastically under the derived ELBO that, when combined with the pre-trained model, is able to represent $p_\theta(x_M|x_O)$ when any subset of data x is missing.

We applied QAVI on a variety of VAEs, including current state-of-the-art hierarchical VAEs, and several datasets. We found that the non-amortized QAVI with Gaussian, Flow, and Mixture posterior approximations outperforms previous heuristic and amortized inference methods developed for the task of data completion with VAEs. Importantly, we find that a Gaussian Mixture posterior is able to capture the multimodality in the missing data distribution well. We believe that our work simply addresses the question of how to apply VAEs to missing data at test time, and hope that such a technique will prove useful when queries are unknown and uncertainty in missing data is high.

References

Burda, Y., Grosse, R., and Salakhutdinov, R. (2015). Importance weighted autoencoders.

Child, R. (2021a). Very deep VAEs generalize autoregressive models and can outperform them on images. *ICLR conference paper, arXiv:2011.10650v2*.

Child, R. (2021b). Very deep {vae}s generalize autoregressive models and can outperform them on images. In *International Conference on Learning Representations*.

Collier, M., Nazabal, A., and Williams, C. (2020). VAEs in the presence of missing data. In *ICML Workshop on the Art of Learning with Missing Values (Artemiss)*.

Cremer, C., Li, X., and Duvenaud, D. (2018). Inference suboptimality in variational autoencoders.

Dechter, R. (2003). *Constraint processing*. Morgan Kaufmann.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Figurnov, M., Mohamed, S., and Mnih, A. (2018). Implicit reparameterization gradients. *Advances in neural information processing systems*, 31.

Gershman, S., Hoffman, M. D., and Blei, D. M. (2012). Nonparametric variational inference. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress.

Ghahramani, Z. and Beal, M. J. (2001). Propagation algorithms for variational Bayesian learning. In *NeurIPS 13*, pages 507–513. MIT Press.

Graves, A. (2016). Stochastic backpropagation through mixture density distributions. *arXiv preprint arXiv:1607.05690*.

Harvey, W., Naderiparizi, S., and Wood, F. (2022). Conditional image generation by conditioning variational autoencoders. In *International Conference on Learning Representations*.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2017). beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347.

Ipsen, N. B., Mattei, P.-A., and Frellsen, J. (2020). notmiwae: Deep generative modelling with missing not at random data. *arXiv preprint arXiv:2006.12871*.

Jaakkola, T. S. and Jordon, M. I. (1999). *Improving the Mean Field Approximation via the Use of Mixture Distributions*, page 163–173. MIT Press, Cambridge, MA, USA.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2):183–233.

Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410.

Kim, Y., Wiseman, S., Miller, A., Sontag, D., and Rush, A. (2018). Semi-amortized variational autoencoders. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2678–2687. PMLR.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. (2016). Improved variational inference with inverse autoregressive flow. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4743–4751, Red Hook, NY, USA. Curran Associates Inc.

Kingma, D. P. and Welling, M. (2014). Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Kingma, D. P. and Welling, M. (2019). An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 12(4):307–392.

Klushyn, A., Chen, N., Kurle, R., Cseke, B., and Smagt, P. v. d. (2019). *Learning Hierarchical Priors in VAEs*. Curran Associates Inc., Red Hook, NY, USA.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Krishnan, R. G., Liang, D., and Hoffman, M. D. (2017). On the challenges of learning with inference networks on sparse, high-dimensional data. *ArXiv*, abs/1710.06085.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017). Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18(1):430–474.

LeCun, Y. and Cortes, C. (2010). MNIST handwritten digit database.

Loaiza-Ganem, G. and Cunningham, J. P. (2019). *The Continuous Bernoulli: Fixing a Pervasive Error in Variational Autoencoders*. Curran Associates Inc., Red Hook, NY, USA.

Marino, J., Yue, Y., and Mandt, S. (2018). Iterative amortized inference. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3403–3412.

Mattei, P.-A. and Frellsen, J. (2018). Leveraging the exact likelihood of deep latent variable models. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Mattei, P.-A. and Frellsen, J. (2019). MIWAE: Deep generative modelling and imputation of incomplete data sets. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4413–4423. PMLR.

Mnih, A. and Gregor, K. (2014). Neural variational inference and learning in belief networks. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, page II–1791–II–1799. JMLR.org.

Nazábal, A., Olmos, P. M., Ghahramani, Z., and Valera, I. (2020). Handling incomplete heterogeneous data using vaes. *Pattern Recognit.*, 107:107501.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. (2011). Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*.

Paisley, J. W., Blei, D. M., and Jordan, M. I. (2012). Variational Bayesian inference with stochastic search. In *International Conference on Machine Learning*.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufman, San Mateo.

Rezende, D. and Mohamed, S. (2015). Variational inference with normalizing flows. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538, Lille, France. PMLR.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286, Bejing, China. PMLR.

Ścibior, A. and Wood, F. (2021). Differentiable particle filtering without modifying the forward pass. *arXiv preprint arXiv:2106.10314*.

Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., and Winther, O. (2016). Ladder Variational Autoencoders. In *Advances in Neural Information Processing Systems*.

Strauss, R. and Oliva, J. (2022). Posterior matching for arbitrary conditioning. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

Vahdat, A. and Kautz, J. (2020a). Nvae: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems*, 33:19667–19679.

Vahdat, A. and Kautz, J. (2020b). NVAE: A deep hierarchical variational autoencoder. In *Neural Information Processing Systems (NeurIPS)*.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1:1–305.

Winn, J. and Bishop, C. M. (2005). Variational message passing. *JMLR*, 6:661–694.

Zagoruyko, S. and Komodakis, N. (2016). Wide residual networks. *ArXiv*, abs/1605.07146.

Zhao, S., Cui, J., Sheng, Y., Dong, Y., Liang, X., Chang, E. I., and Xu, Y. (2021). Large scale image completion via co-modulated generative adversarial networks. In *International Conference on Learning Representations (ICLR)*.