Variational Selective Autoencoder: Learning from Partially-Observed Heterogeneous Data

Yu Gong[1,2] **Hossein Hajimirsadeghi**[1] **Jiawei He**[1] **Thibaut Durand**[1] **Greg Mori**[1,2]
[1]Borealis AI [2]Simon Fraser University

Abstract

Learning from heterogeneous data poses challenges such as combining data from various sources and of different types. Meanwhile, heterogeneous data are often associated with missingness in real-world applications due to heterogeneity and noise of input sources. In this work, we propose the variational selective autoencoder (VSAE), a general framework to learn representations from partially-observed heterogeneous data. VSAE learns the latent dependencies in heterogeneous data by modeling the joint distribution of observed data, unobserved data, and the imputation mask which represents how the data are missing. It results in a unified model for various downstream tasks including data generation and imputation. Evaluation on both low-dimensional and high-dimensional heterogeneous datasets for these two tasks shows improvement over state-of-the-art models.

1 Introduction

Learning from data is an integral part of artificial intelligence. A typical assumption of learning algorithms that the data is fully-observed is clearly unrealistic in many settings, therefore handling missing data has been a long-standing problem (Ghahramani and Jordan, 1994; Schafer, 1997). The data generation process combined with the observation mechanism by which data are hidden makes learning from those data much more complex. Moreover, conventional algorithms rely heavily on clean *homogeneous* data, yet varied, *heterogeneous* data are a common setting for learning. In fact, heterogeneity is ubiquitous in a variety of platforms from healthcare to finance to social networks to manufacturing systems (He, 2017). By the nature of the observation mechanism of the data, partial observability is often associated with heterogeneity. For example, a bank client is more likely to conceal the income amount than the attributes like gender or age; a doctor can never perform all medical tests for a patient, but choose to collect test results based on particular symptoms and expertise. In this work we present a deep latent variable model for representation learning from partially-observed heterogeneous data.

Deep generative models have been shown to be effective in a variety of homogeneous data learning (Bando et al., 2018; Gulrajani et al., 2017; Mescheder et al., 2017; Yang et al., 2017). However, learning these models from heterogeneous data introduces new challenges. In particular, our definition of heterogeneity spans a wide range of forms from data type to data source/modality. For example, we may have categorical or numerical data from different distributions, or mixed modalities representing images, text, and audio. The main challenge is how to align and integrate heterogeneous data to model the joint distribution. Our proposed latent variable model handles this effectively by selecting appropriate proposal distributions, and performing the integration in a latent space instead of the input space.

Learning from partially-observed data is another challenge in deep generative models. Naive solutions such as ignoring or zero-imputing missing data will likely degrade performance by introducing sparsity bias (Yi et al., 2019). Having a model designed to learn from incomplete data not only increases the application spectrum of deep learning algorithms but also benefits downstream tasks such as data imputation, which remains an open and challenging area of research.

Some prior work requires fully-observed data for training (Suzuki et al., 2016; Ivanov et al., 2019), or makes the assumption that data are missing completely at random (MCAR) (Yoon et al., 2018; Li et al., 2019), which assumes *missingness* (the manner in which data are missing) occurs independently from the data. Our method relaxes these assumptions by learning the joint distribution of data and missingness patterns (or *mask*).

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

In this work, we propose the variational selective autoencoder (VSAE), a general and flexible model for representation learning from partially-observed heterogeneous data. The proposed deep latent variable model is capable of capturing hidden dependencies within partially-observed heterogeneous data by performing selection and integration in the latent representation. It learns the joint distribution of data/mask without strong assumptions about missingness mechanism, resulting in applications for data generation and imputation. In particular, it can be trained effectively with a single objective to impute missing data from any combination of observed data. Extensive evaluation on challenging low-dimensional and high-dimensional heterogeneous data shows improvement over state-of-the-art models. The contributions are summarized as follows:

- A novel selective proposal distribution efficiently learns representations from partially-observed heterogeneous data.
- The proposed method models the joint distribution of the data and the imputation mask, resulting in a unified model for various tasks including generation and imputation.
- VSAE does not make restrictive assumptions on the missingness mechanism, expanding the scope of scenarios in which data imputation can be effectively learned.

2 Related Work

Learning from Heterogeneous Data. Most existing approaches modeling statistical dependencies in unstructured heterogeneous data focus on obtaining alignments and subsequently modeling relationships between different domains (Kim et al., 2017; Zhu et al., 2017; Castrejón et al., 2016). However, there has been little progress in learning the joint distribution of the full data comprising different domains. Other methods (Misra et al., 2016; Zhang et al., 2018) handle heterogeneity in labels or datasets in weakly-supervised learning settings. MVAE (Wu and Goodman, 2018) uses a product-of-experts inference network to solve the inference problem in multi-model setting. Our work focuses on modeling all types of heterogeneity and demonstrates effectiveness in data generation and imputation applications.

Learning from Partially-Observed Data. Classical methods dealing with missing data such as MICE (Buuren and Groothuis-Oudshoorn, 2010) and MissForest (Stekhoven and Bühlmann, 2011) typically learn discriminative models to impute missing features. Advanced by deep neural networks, several models have also been developed to address data imputation based on autoencoders (Gondara and Wang, 2018; Vincent et al., 2008), generative adversarial networks (GANs) (Li et al., 2019; Yoon et al., 2018), and autoregressive models (Bachman and Precup, 2015). In this work, we focus on improving deep latent variable models to efficiently learn from partially-observed heterogeneous data. Deep latent variable models (DLVMs) are generative models that can map complex raw input to a flexible latent representation and have recently gained attention on handling partially-observed data due to the flexibility of generative modeling and representation learning. To tackle the intractable posterior of DLVMs, variational autoencoder (VAE) first uses deep neural networks to approximate the posterior and maximizes a variational evidence lower bound (ELBO). Based on VAE, prior work (Ivanov et al., 2019; Ma et al., 2019; Mattei and Frellsen, 2019; Nazabal et al., 2020; Collier et al., 2020) attempted to improve DLVMs under strong missingness mechanism assumptions. VAEAC (Ivanov et al., 2019) imputed attributes conditional on observed ones by learning from fully-observed data; Partial VAE (Ma et al., 2019) encoded observed data with a permutation invariant set function; MIWAE (Mattei and Frellsen, 2019) introduced a tighter bound using importance sampling under a relaxed Missing At Random (MAR, refer to Sec. 3) assumption; HI-VAE (Nazabal et al., 2020) factored decoders on low-dimensional heterogeneous data under MCAR. These methods jointly map the sources (a.k.a. *attribute*) of heterogeneous data into a holistic latent space, which adds unavoidable noise to the latent space as the distribution of one attribute can be far from the others.

3 Background

Problem Statement. We represent any *heterogeneous* data point as a set of random variables $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2 ..., \mathbf{x}_M]$ representing different *attributes* collected from multiple sources. The type and size of each attribute $\mathbf{x}_i$ can vary. It can be either high-dimensional (e.g. multimedia data) or low-dimensional (e.g. tabular data). We define an M-dimensional binary *mask* variable $\mathbf{m} \in \{0, 1\}^M$ to represent the missingness: for the i-th attribute, $m_i = 1$ if it is observed and 0 otherwise. Thus we can induce *observed attributes* by the set $\mathbb{O} = \{i | m_i = 1\}$ and *unobserved attributes* by the complementary set $\mathbb{U} = \{i | m_i = 0\}$. Accordingly, we denote the collective representation of the observed attribute with $\mathbf{x_o} = [\mathbf{x}_i | m_i = 1]$ and unobserved attributes with $\mathbf{x_u} = [\mathbf{x}_i | m_i = 0]$. In general, every instance has a different set of $\mathbf{x_o}$ as well as $\mathbf{x_u}$, determined by the mask variable $\mathbf{m}$.

Missingness Mechanism. The generative process of incomplete data can be modeled by the joint distribution $p(\mathbf{x_o}, \mathbf{x_u}, \mathbf{m})$. Little and Rubin (2019) categorize the missingness mechanism into three types based on the dependence between the data and mask as follows,

Missing Completely At Random (MCAR). Missingness is completely independent of data,

$$p(\mathbf{x_o}, \mathbf{x_u}, \mathbf{m}) = p(\mathbf{x_o}, \mathbf{x_u})p(\mathbf{m}) \tag{1}$$

Missing At Random (MAR). Missingness depends only on observed attributes,

$$p(\mathbf{x_o}, \mathbf{x_u}, \mathbf{m}) = p(\mathbf{x_o}, \mathbf{x_u})p(\mathbf{m}|\mathbf{x_o}) \tag{2}$$

Not Missing At Random (NMAR). Missingness depends on both observed and unobserved attributes,

$$p(\mathbf{x_o}, \mathbf{x_u}, \mathbf{m}) = p(\mathbf{x_o}, \mathbf{x_u})p(\mathbf{m}|\mathbf{x_o}, \mathbf{x_u}) \tag{3}$$

Most prior work on learning from partially-observed data follows the MCAR or MAR assumption since the factorization in Eq. (1) and Eq. (2) decouples the mask $\mathbf{m}$ from $\mathbf{x_u}$ in the integral of the likelihood function. A common approach is to define the log-likelihood of an incomplete dataset by marginalizing over the unobserved attributes $\log[\int p(\mathbf{x_o}, \mathbf{x_u}, \mathbf{m}) d\mathbf{x_u}] = \log p(\mathbf{x_o}, \mathbf{m})$ and ignoring the underlying missing mechanism. Our goal is to go beyond this simple but restrictive solution and model the joint distribution of the data and the mask.

4 Proposed Method

Variational Selective Autoencoder (VSAE) aims to relax the strong missingness assumption and model the joint distribution $p(\mathbf{x}, \mathbf{m}) = \int p(\mathbf{x}, \mathbf{m}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$. The latent variables $\mathbf{z}$ are inferred from data instances with the mask to capture the dependence among the attributes and the mask. The data and missingness information embedded in the latent variables can together guide data imputation and generation. At a high level, as illustrated in Fig. 1, VSAE is formulated with individual encoders for observed attributes and a collective encoder for unobserved ones to construct the selective proposal distribution. The selected latent variables of each attribute are aggregated and decoded to reconstruct the mask and all attributes independently. To tackle the intractability of unobserved attributes $\mathbf{x_u}$ and latent variables $\mathbf{z}$, we formulate the learning procedure as an external expectation maximization (EM) nested with variational inference (VI). In the following, we provide details on the VSAE model.

4.1 Model description

Following the VAE[1] (Kingma and Welling, 2014; Rezende et al., 2014), we construct a proposal distribution $q(\mathbf{z}|\mathbf{x}, \mathbf{m})$ to approximate the intractable true posterior. With the inclusion of the novel selective proposal distribution, we expand the parameters of inference networks to $\{\boldsymbol{\phi}, \boldsymbol{\psi}\}$, where $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ represent encoder parameter for observed and unobserved attributes. Following the same fashion, the parameters of generative networks are expanded to $\{\boldsymbol{\theta}, \boldsymbol{\epsilon}\}$, with $\boldsymbol{\theta}$ denoting the decoder parameter for the data, and $\boldsymbol{\epsilon}$ for the mask. The variational evidence lower bound of $\log p(\mathbf{x}, \mathbf{m})$ can thus be derived as

$$\begin{aligned}\mathcal{L}_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\theta},\boldsymbol{\epsilon}}(\mathbf{x}, \mathbf{m}) = &\underbrace{\mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{z}|\mathbf{x},\mathbf{m})}[\log p_{\boldsymbol{\theta},\boldsymbol{\epsilon}}(\mathbf{x}, \mathbf{m}|\mathbf{z})]}_{\text{cond. LL}} \\ &- \underbrace{D_{\mathrm{KL}}(q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{z}|\mathbf{x}, \mathbf{m})||p(\mathbf{z}))}_{\text{KL Regularizer}},\end{aligned} \tag{4}$$

Factorized Latent Space. We assume the latent space can be factorized w.r.t. M attributes,

$$p(\mathbf{z}) = \prod_{i=1}^{M} p(\mathbf{z}_i), \quad q(\mathbf{z}|\mathbf{x}, \mathbf{m}) = \prod_{i=1}^{M} q(\mathbf{z}_i|\mathbf{x}, \mathbf{m}) \tag{5}$$

Priors $p(\mathbf{z}_i)$ are standard Gaussians and proposal distributions $q(\mathbf{z}_i|\mathbf{x}, \mathbf{m})$ are Gaussians with inferred means and diagonal covariances. This factorization separates the encoding of each attribute and efficiently yields a distribution for latent variables by assuming the latent variables are conditionally independent given the data and mask. Hence, it provides a mechanism to decouple the heterogeneity in the raw data space while integrating them efficiently in the latent space.

Selective Proposal Distribution. The standard proposal distribution of VAEs, inferred from fully-observed data, is not applicable for partially-observed input. To circumvent this, we introduce our *selective proposal distribution* for each latent variable:

$$q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{z}_i|\mathbf{x}, \mathbf{m}) = \begin{cases} q_{\boldsymbol{\phi}}(\mathbf{z}_i|\mathbf{x}_i) & \text{if } m_i = 1 \\ q_{\boldsymbol{\psi}}(\mathbf{z}_i|\mathbf{x_o}, \mathbf{m}) & \text{if } m_i = 0 \end{cases} \tag{6}$$

This conditional selection of proposal distribution is determined by the mask variable. Accordingly, we subdivide the inference network into two types:

Attributive Proposal Network. $q_{\boldsymbol{\phi}}(\mathbf{z}_i|\mathbf{x}_i)$, inferred merely from the individual observed attribute and selected for an observed attribute;

Collective Proposal Network. $q_{\boldsymbol{\psi}}(\mathbf{z}_i|\mathbf{x_o}, \mathbf{m})$, collecting all observed values and the mask to produce the proposal distribution and selected for an unobserved attribute. This formulation aids VAE encoders by explicitly focusing on the relevant inputs and ignoring the less informative ones.

Latent Variable Aggregation. We sample the latent variables for all attributes using Eq. (6). Next, to

[1]Refer to Appendix A for a detailed description of VAE.

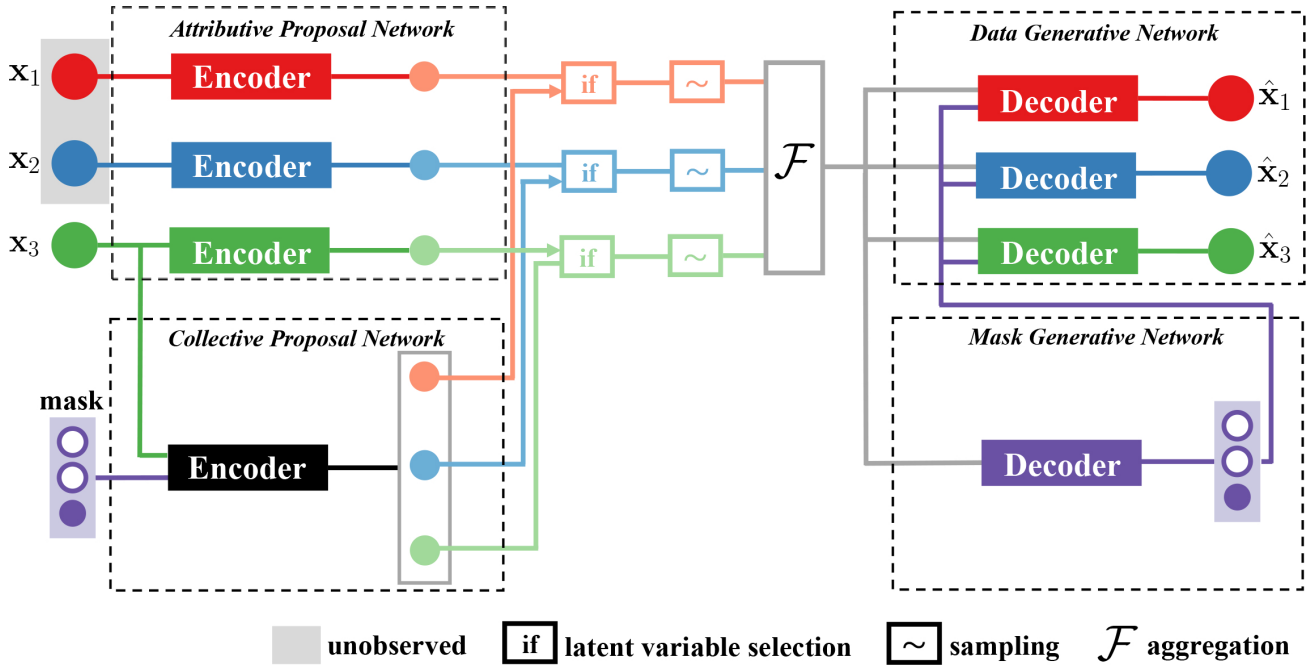

Figure 1: **Model overview.** The input includes data *attributes* (denoted by different colors — $\mathbf{x}_1, \mathbf{x}_2$ are unobserved; $\mathbf{x}_3$ is observed) and *mask*. The ***attributive proposal network*** takes the attributes as input, while the ***collective proposal network*** takes the observed attributes and the mask as input. We use if-condition to denote the selective proposal distribution (Eq. 6, the arrowed one is selected). The aggregated latent codes are fed to both ***mask generative network*** and ***data generative network***. The output of the mask generative network will be fed to each decoder of the data generative network as extra condition. Standard normal prior is not plotted for simplicity. All components are trained simultaneously in an end-to-end manner.

capture the dependencies between observed attributes, unobserved attributes and mask, an aggregation function $\mathcal{F}(\cdot)$ is performed before the decoders. We use *concatenation* as $\mathcal{F}(\cdot)$, though it can be any aggregation function in general. The conventional VAEs, however, often aggregate the attributes naively in the raw data space. Consequently, the heterogeneity and partially-observed nature will restrain those models from learning informative representations.

Data & Mask Generative Networks. By applying the chain rule, the conditional log-likelihood $\log p_{\boldsymbol{\theta},\boldsymbol{\epsilon}}(\mathbf{x},\mathbf{m}|\mathbf{z})$ in Eq. (4) is decomposed as mask conditional log-likelihood $\log p_{\boldsymbol{\epsilon}}(\mathbf{m}|\mathbf{z})$ and data conditional log-likelihood $\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{m},\mathbf{z})$. The mask and data are reconstructed from shared $\mathbf{z}$ through the ***mask generative network*** and ***data generative network*** shown in Fig. 1. Further, the data conditional log-likelihood factorizes over the attributes assuming the reconstructions are conditionally independent given $\mathbf{m}$ and $\mathbf{z}$:

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{m},\mathbf{z}) = \underbrace{\sum_{i\in\mathbb{O}} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i|\mathbf{m},\mathbf{z})}_{\text{Observed}} + \underbrace{\sum_{j\in\mathbb{U}} \log p_{\boldsymbol{\theta}}(\mathbf{x}_j|\mathbf{m},\mathbf{z})}_{\text{Unobserved}} \quad (7)$$

Expectation Maximization. The ELBO (Eq. (4)) is hard to maximize since $\mathbf{x_u}$ is unobserved. We can use expectation maximization (EM) algorithm (Dempster et al., 1977) to handle its intractability (refer to Appendix B for full derivation). EM alternates between inferring the unobserved data given the parameters (E step) and optimizing the parameters given the "filled in" data (M step).

- E **step:** By taking an expectation over $\mathbf{x_u}$,

$$\mathcal{L}'_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\theta},\boldsymbol{\epsilon}}(\mathbf{x_o},\mathbf{m}) = \mathbb{E}_{\mathbf{x_u}}[\mathcal{L}_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\theta},\boldsymbol{\epsilon}}(\mathbf{x_o},\mathbf{x_u},\mathbf{m})] \quad (8)$$

We now expand Eq. (8) into conditional log-likelihood and KL divergence terms. Since only the unobserved attributes conditional log-likelihood depends on $\mathbf{x_u}$, with $\mathbf{z}_i \sim q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{z}_i|\mathbf{x},\mathbf{m})$ given by Eq. (6), we obtain

$$\mathcal{L}'_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\theta},\boldsymbol{\epsilon}}(\mathbf{x_o},\mathbf{m}) = \underbrace{\mathbb{E}_{\mathbf{z}}\big[\sum_{j\in\mathbb{U}} \mathbb{E}_{\mathbf{x}_j}[\log p_{\boldsymbol{\theta}}(\mathbf{x}_j|\mathbf{m},\mathbf{z})]\big]}_{\text{Unobserved attributes cond. LL}} +$$

$$\underbrace{\mathbb{E}_{\mathbf{z}}\big[\sum_{i\in\mathbb{O}} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i|\mathbf{m},\mathbf{z})\big]}_{\text{Observed attributes cond. LL}} + \underbrace{\mathbb{E}_{\mathbf{z}}[\log p_{\boldsymbol{\epsilon}}(\mathbf{m}|\mathbf{z})]}_{\text{Mask cond. LL}} -$$

$$\underbrace{\sum_{i=1}^{M} \mathbb{E}_{\mathbf{z}_i}[\log q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{z}_i|\mathbf{x},\mathbf{m}) - \log p(\mathbf{z}_i)]}_{\text{KL Regularizer}} \quad (9)$$

In Eq. (9), direct calculation of unobserved attributes conditional log-likelihood is intractable. Instead, during training given the learned generative model $p(\mathbf{x_o}, \mathbf{x_u}, \mathbf{m}, \mathbf{z})$ with VI, we can generate conditional distribution of unobserved attribute $\mathbf{x}_j$ from:

$$\hat{p}(\mathbf{x}_j|\mathbf{x_o}, \mathbf{m}) = \int \Phi(\mathbf{z}) p_{\boldsymbol{\theta}^*}(\mathbf{x}_j|\mathbf{m}, \mathbf{z}) d\mathbf{z} \tag{10}$$

where $\Phi(\mathbf{z}) := \prod_{i\in\mathbb{O}} q_{\boldsymbol{\phi}^*}(\mathbf{z}_i|\mathbf{x}_i) \prod_{j\in\mathbb{U}} p(\mathbf{z}_i)$ and $*$ denotes the parameters learned up to the current iteration. It enables us to estimate each unobserved terms with $\mathbb{E}_{\mathbf{x}_j\sim\hat{p}(\mathbf{x}_j|\mathbf{x_o},\mathbf{m})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}_j|\mathbf{m}, \mathbf{z})]$.

• M **step:** We can therefore maximize the final objective function (Eq. (9)) with $\mathbf{x_u}$ sampled from Eq. (10). Conventional methods of learning from partially-observed data simply marginalizes and maximizes $\log \int p(\mathbf{x_o}, \mathbf{x_u}, \mathbf{m}) d\mathbf{x_u} = \log p(\mathbf{x_o}, \mathbf{m})$ by ignoring $\mathbf{x_u}$. Our procedure can be viewed as maximizing the expected lower bound $\mathbb{E}_{\mathbf{x_u}}[\mathcal{L}_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\theta},\boldsymbol{\epsilon}}(\mathbf{x_o}, \mathbf{x_u}, \mathbf{m})]$ to approximately maximize $\mathbb{E}_{\mathbf{x_u}}[\log p(\mathbf{x_o}, \mathbf{x_u}, \mathbf{m})]$.

Empirically, given a partially-observed batch, our two-stage scheme is: (1) decode the latent codes drawn from the prior $p(\mathbf{z}_i)$ of unobserved attributes and the attributive proposal distribution $q_{\boldsymbol{\phi}^*}(\mathbf{z}_i|\mathbf{x}_i)$ of observed attributes to generate $\mathbf{x_u}$ by the decoders learned so far; (2) re-input the same incomplete batch to calculate all observed terms and estimate the unobserved term with Eq. (10), then optimize over all parameters. We generate 100 samples of $\mathbf{x_u}$ to take expectation and experiments show that it gives an effective estimation to the full expectation.

Another possible alternative in E step is to infer the latent variables of unobserved attributes from the collective proposal distribution $q_{\boldsymbol{\psi}^*}(\mathbf{z}|\mathbf{x_o}, \mathbf{m})$ but empirically we find the prior performs better. The possible explanation lies in two-fold: (i) Approximating the true posterior $p(\mathbf{z}|\mathbf{x_u}, \mathbf{x_o}, \mathbf{m})$ is challenging without $\mathbf{x_u}$. At early stages of training, the parametric proposal distribution is not learned well to match the posterior. Hence, the parameter-free prior is a more stable choice for the noisy unobserved attributes; (ii) If sampling from the prior for unobserved attributes, the aggregated latent variables are still conditioned on $\mathbf{x_o}$ from the attributive proposal distribution. Compared to collective proposal distribution $q_{\boldsymbol{\psi}^*}(\mathbf{z}|\mathbf{x_o}, \mathbf{m})$ for $\mathbf{x_u}$, sampling from the prior does not depend on $\mathbf{m}$. Since both data and mask generative processes have not yet been accurately captured, it can encourage focus on generating $\mathbf{x_u}$.

4.2 Model applications

Unlike conventional data imputation models, the generative model $p(\mathbf{x}, \mathbf{m}, \mathbf{z})$ is learned. Therefore, VSAE constructs a unified framework for data imputation, data generation and mask generation.

Data Imputation. The aim is to impute the missing data given the observed data, thus it can be viewed as conditional generation of the unobserved attributes. This can be performed by sampling the latent codes for all attributes using $q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{z}_i|\mathbf{x}, \mathbf{m})$ in Eq. (6). Next, the aggregated latent codes and mask are given to the decoders of the unobserved attributes for generation. This process can be described as $p(\mathbf{x_u}|\mathbf{x_o}, \mathbf{m}) \approx \int p_{\boldsymbol{\theta}}(\mathbf{x_u}|\mathbf{m}, \mathbf{z}) q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\mathbf{z}|\mathbf{x_o}, \mathbf{m}) d\mathbf{z}$.

Data & Mask Generation. Given random samples from a standard Gaussian prior $p(\mathbf{z})$, we can generate masks using the mask generative network. Next, the sampled latent codes and the generated mask are given to the data generative network to generate data.

Modeling the mask distribution in our framework not only helps to inform the data generation of the relevant missingness mechanism but also enables to sample the mask itself. MisGAN (Li et al., 2019) also employs a mask generation process. However, MisGAN generates the mask assuming the mask is MCAR, whereas VSAE generates the mask assuming the mask is NMAR. Mask generation may have applications like synthesizing incomplete data or obtaining the potential mask if the real mask is not available (e.g. if data is corrupted rather than only missing).

4.3 Scalability

The complexity grows linearly w.r.t the number of attributes M and their dimensions. In practice, heterogeneous data are often either *tabular* (large M but low-dimensional) or *multi-modal* (high-dimensional but small M), thus practically tractable. For large M with high-dimensional attributes (e.g. URLs), one may use embedding layers to densify these categorical attributes to low-dimensional vectors, or manually group subset of the attributes to constrain the M as a feasible constant.

5 Experiments

5.1 Tabular Data

Tabular data are ordered arrangements of rows and columns. Each row is a sample with multiple attributes (typically low-dimensional) and each column is a single attribute collected heterogeneously. Due to communication or privacy issues, those attributes of data samples often are partially-observed. For this scenario, we choose **UCI repository** which contains various tabular datasets of numerical or categorical attributes. In all experiments, min-max normalization is applied to pre-process the numerical data and the unobserved

	Phishing	Mushroom	Yeast	Whitewine	Heart (mixed)	
Attribute type	categorical	categorical	numerical	numerical	categorical	numerical
AE	0.348 ± 0.004	0.556 ± 0.009	0.737 ± 0.035	$0.3772 \pm \Delta$	0.514 ± 0.026	0.650 ± 0.006
VAE	$0.274 \pm \Delta$	0.470 ± 0.016	$0.461 \pm \Delta$	$0.3724 \pm \Delta$	$\mathbf{0.485 \pm 0.016}$	0.630 ± 0.025
CVAE w/ mask	$0.241 \pm \Delta$	$0.445 \pm \Delta$	$0.445 \pm \Delta$	$0.3726 \pm \Delta$	0.516 ± 0.015	0.605 ± 0.005
MVAE	0.308 ± 0.016	0.586 ± 0.017	0.442 ± 0.016	$0.3732 \pm \Delta$	0.517 ± 0.010	0.636 ± 0.009
HI-VAE	$0.238 \pm \Delta$	0.470 ± 0.008	0.429 ± 0.005	$\mathbf{0.3720 \pm \Delta}$	$\mathbf{0.481 \pm 0.012}$	0.603 ± 0.008
VSAE (ours)	$\mathbf{0.237 \pm \Delta}$	$\mathbf{0.416 \pm 0.009}$	$\mathbf{0.419 \pm 0.008}$	$\mathbf{0.3719 \pm \Delta}$	$\mathbf{0.482 \pm 0.014}$	$\mathbf{0.579 \pm 0.015}$

Table 1: **MCAR Data Imputation on UCI datasets.** We consider categorical, numerical and mixed tabular datasets in the missing ratio of 0.5. Categorical and numerical attributes are evaluated by **PFC** and **NRMSE** respectively, lower is better for both. We show mean and standard deviation over 3 independent trials. $\Delta < 0.0005$.

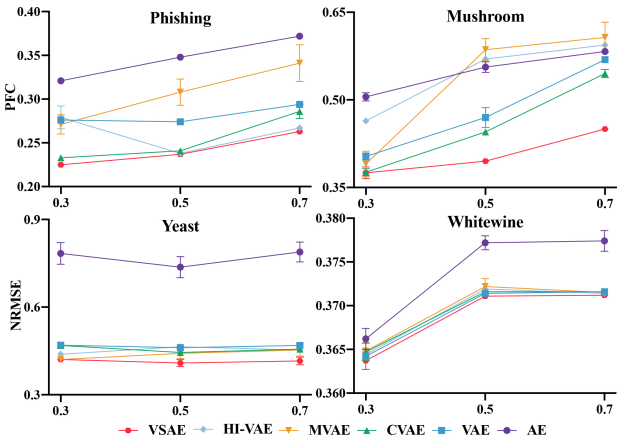

Figure 2: **Data Imputation on UCI datasets.** X-axis is missing ratio. Categorical (top)/numerical (bottom) datasets are evaluated by **PFC**/**NRMSE** with mean/std over 3 independent trials. Lower is better.

dimensions are replaced by standard normal noise. We split the training and test set with size ratio 4:1 and use 20% of training data as a validation set to choose the best model. Mean-squared error, cross-entropy and binary cross-entropy are used as reconstruction loss for numerical, categorical and mask variables, respectively.

Data Imputation. We first consider a data imputation experiment—imputing unobserved attributes given observed attributes and mask. VSAE can be used in this case as in Sec. 4.2. We report the standard measures: **NRMSE** (RMSE normalized by the standard deviation of ground truth features, averaged over all features) and **PFC** (proportion of falsely classified attributes of each feature, averaged over all features) for numerical and categorical attributes. The evaluation is under various missing mechanisms by synthesizing masks following different rules.

MCAR masking. We randomly sample from independent Bernoulli distributions with predefined missing ratios to mimic MCAR missing mechanism. VSAE is

	Method	MAR	NMAR
Yeast	MIWAE	$\mathbf{0.475 \pm 0.005}$	0.456 ± 0.036
	VSAE (ours)	$\mathbf{0.472 \pm 0.006}$	$\mathbf{0.425 \pm 0.007}$
Whitewine	MIWAE	$0.3834 \pm \Delta$	$0.3723 \pm \Delta$
	VSAE (ours)	$\mathbf{0.3825 \pm \Delta}$	$\mathbf{0.3717 \pm \Delta}$

Table 2: **Non-MCAR Data Imputation.** We show mean and standard deviation of **NRMSE** over 3 independent trials, lower is better. $\Delta < 0.0005$.

compared with the deterministic autoencoder (AE), VAE, conditional VAE (CVAE) (Sohn et al., 2015) conditioned on the mask, multi-modal MVAE (Wu and Goodman, 2018) and HI-VAE (Nazabal et al., 2020). We use publicly released codes of MVAE/HI-VAE, and implement other baselines with the same backbone and at least as many parameters as VSAE.

Table 1 shows that VSAE outperforms other methods on all datasets under MCAR missing mechanism in the missing ratio of 0.5. Fig. 2 illustrates VSAE generally achieves lower error along with relatively lower variance in all missing ratios. When the missing ratio increases (i.e. more data attributes become unobserved), VSAE consistantly maintains stable performance on most of the datasets. Conversely, we can observe a performance drop along with higher variance in the case of baselines. As the missing ratio increases, the attributive proposal network of VSAE maintains the same input, while the encoders of other methods have to learn to focus on the valuable information. We believe the selection mechanism of proposal distribution in VSAE mitigates this negative effect.

In Fig. 2, a seemingly strange behaviour is observed — some models (e.g. HI-VAE for *Phishing*) perform better with more missing data. We hypothesize that: (i) The missingness provides a natural way of dropout on the raw data space. Under certain missing ratios, it possibly reduces overfitting. Hence, it can help the model to generalize; (ii) To synthesize the missing mechanism, we sample and fix a mask on each dataset

	FashionMNIST + label (PFC)		MNIST + MNIST		CMU-MOSI		
Attribute type	Image	Label	Digit–1	Digit–2	Text	Audio	Image
AE	0.1105 ± 0.001	0.366 ± 0.01	$0.1077 \pm \Delta$	$0.1070 \pm \Delta$	0.035 ± 0.003	0.224 ± 0.025	0.019 ± 0.003
VAE	$0.0885 \pm \Delta$	0.411 ± 0.01	$0.0734 \pm \Delta$	$0.0682 \pm \Delta$	$0.034 \pm \Delta$	0.202 ± 0.003	$\mathbf{0.017 \pm \Delta}$
CVAE w/ mask	$0.0887 \pm \Delta$	0.412 ± 0.01	$0.0733 \pm \Delta$	$0.0679 \pm \Delta$	$0.043 \pm \Delta$	0.257 ± 0.002	$0.020 \pm \Delta$
MVAE	0.1402 ± 0.026	0.374 ± 0.07	$0.0760 \pm \Delta$	$0.0802 \pm \Delta$	$0.44 \pm \Delta$	0.213 ± 0.001	$0.025 \pm \Delta$
HI-VAE	0.1575 ± 0.006	0.405 ± 0.01	$0.0772 \pm \Delta$	$0.0725 \pm \Delta$	$0.047 \pm \Delta$	0.211 ± 0.005	$0.0267 \pm \Delta$
VSAE (ours)	$\mathbf{0.0874 \pm \Delta}$	$\mathbf{0.356 \pm 0.01}$	$\mathbf{0.0712 \pm \Delta}$	$\mathbf{0.0663 \pm \Delta}$	$\mathbf{0.033 \pm \Delta}$	$\mathbf{0.200 \pm \Delta}$	$\mathbf{0.017 \pm \Delta}$

Table 3: **Data Imputation on multi-modal datasets.** Missing ratio is 0.5. We evaluate each dataset w.r.t. each attribute—label attribute is evaluated by **PFC**, image attributes of MNIST and FashionMNIST are evaluated by **MSE** averaged over pixels, other attributes are evaluated by **MSE**. Lower is better for all. We show mean and standard deviation over 3 independent trials. $\Delta < 0.001$.

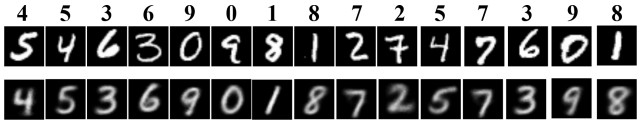

Figure 3: **Data Imputation on MNIST+MNIST.** Top is the labels of unobserved digit via pre-defined rules; middle is the observed attribute; bottom shows the imputation of unobserved attribute from VSAE.

Figure 4: **Data Generation on MNIST+MNIST.** Images are generated without conditional information. As shown, the correspondence between attributes are preserved in stochastic data generation.

for all models, but the model itself may also be sensitive to the mask randomness. Under a certain random sampling of the mask, some corrupted data or outliers, where one model is sensitive, might be observed and degrade the performance of this model. For example, on *Phishing* dataset, HI-VAE performs slightly better at 0.5 than 0.3). The error bar is wider at 0.3, but relatively narrow at 0.5 and 0.7. Some outliers observed at 0.3 may contribute to this variance. It performs worse again at 0.7, as the positive effect by masking out the outliers is not dominant when the missing ratio is too high.

Non-MCAR masking. MIWAE (Mattei and Frellsen, 2019) assumes data are MAR and approximates the conditional expectation of the unobserved attributes using importance sampling. We follow it to mimic non-MCAR missing mechanisms on UCI numerical datasets — MAR: 25% attributes are default observed, then sample the remaining mask from sigmoid($\frac{1}{M}\sum_{k=1}^{K}\mathbf{x}_k$), where K is the number of observed attributes; NMAR: sample the mask for $\mathbf{x}_i$ from sigmoid($\mathbf{x}_i$). We use the public code of MIWAE and the same imputation estimates (single imputation of 1000 importance samplings) for fair comparison. Table 2 indicates VSAE can outperform the state-of-the-art non-MCAR model MIWAE in the non-MCAR setting, particularly for the NMAR, as VSAE models the joint distribution of attributes and mask without introducing the presumably false independence among them.

Mask Generation. VSAE enables us to generate data and mask from the learned generative model $p_{\boldsymbol{\theta},\boldsymbol{\epsilon}}(\mathbf{x},\mathbf{m},\mathbf{z})$. We show mask generation results on UCI and data generation on multi-modal datasets (Sec. 5.2), since the sole data generation is not qualitatively or quantitatively measurable in the case of UCI datasets. After training, we can sample from the prior to generate the mask on the corresponding dataset.

We evaluate mask generation by calculating the average proportion of missing attributes ($m_i = 0$) on 100 sampled masks. In the **MCAR** setting, averaged over datasets, we get 0.312±0.016, 0.496±0.009, 0.692±0.005 for the ground truth missing ratios of 0.3, 0.5, 0.7. In the **MAR** setting where the mask depends on the dataset, we get 0.631±0.026 and 0.665±0.019 for the ground truth missing ratios of 0.626 (Whitewine) and 0.676 (Yeast). Similarly, in the **NMAR** setting we get 0.513±0.021 and 0.592±0.013 for the ground truth missing ratios of 0.502 (Whitewine) and 0.583 (Yeast). It indicates VSAE is capable of accurately learning the mask distribution. Interestingly, we find improvement in the performance by conditioning the reconstructed mask variable on the data decoders. We speculate that this may be because the learned distributional mask variable can inform the data decoder of the missingness distributed in the data space, which in turn allows the potential missing mechanism to guide the data generative process.

	MCAR		MAR		NMAR	
	Yeast	**Whitewine**	**Yeast**	**Whitewine**	**Yeast**	**Whitewine**
Missing ratio	0.5	0.5	0.676	0.626	0.583	0.502
VSAE (collective only)	0.426 ± 0.001	$0.3730 \pm \Delta$	$0.480 \pm \Delta$	$0.3837 \pm \Delta$	$\mathbf{0.427 \pm 0.005}$	$0.3737 \pm \Delta$
VSAE (attributive only)	0.457 ± 0.017	$0.3728 \pm \Delta$	0.493 ± 0.005	$0.3827 \pm \Delta$	0.434 ± 0.006	$0.3729 \pm \Delta$
VSAE (w/o mask modeling)	$\mathbf{0.418 \pm 0.004}$	$0.3720 \pm \Delta$	0.485 ± 0.006	$0.3834 \pm \Delta$	$0.437 \pm \Delta$	$0.3724 \pm \Delta$
No EM	$\mathbf{0.426 \pm 0.012}$	$0.3723 \pm \Delta$	0.476 ± 0.001	$\mathbf{0.3821 \pm \Delta}$	0.432 ± 0.003	$0.3727 \pm \Delta$
EM with 1 sampling	0.433 ± 0.002	$\mathbf{0.3719 \pm \Delta}$	$\mathbf{0.473 \pm \Delta}$	$0.3825 \pm \Delta$	$\mathbf{0.426 \pm 0.003}$	$0.3721 \pm \Delta$
EM with 1000 samplings	$\mathbf{0.418 \pm 0.008}$	$\mathbf{0.3718 \pm \Delta}$	$\mathbf{0.471 \pm 0.007}$	$0.3826 \pm \Delta$	$\mathbf{0.424 \pm 0.005}$	$\mathbf{0.3716 \pm \Delta}$
VSAE (ours)	$\mathbf{0.419 \pm 0.008}$	$\mathbf{0.3719 \pm \Delta}$	$\mathbf{0.472 \pm 0.006}$	$0.3825 \pm \Delta$	$\mathbf{0.425 \pm 0.007}$	$\mathbf{0.3717 \pm \Delta}$

Table 4: **Ablation study on modules of VSAE.** We consider experiments on *Yeast* and *Whitewine*. The masks of MCAR, MAR and NMAR are synthesized by the rules mentioned in Sec. 5.1. We evaluate each dataset by **NRMSE**. Lower is better. We show mean and standard deviation over 3 independent trials. $\Delta < 0.0005$.

5.2 Multi-modal Data

Baltrušaitis et al. (2017) defined *multi-modal data* as data of multiple *modalities*, where each modality is a way to sense the world—seeing, hearing, feeling, etc. However, our definition of *multi-modal* covers a wider spectrum where the data could be of the same type (e.g. image) but from different distributions (e.g. different shapes). In the manner multi-modal data are collected or represented, we can safely treat multi-modal data (typically high-dimensional) as a type of heterogeneous data. In the following, we use *attribute* and *modality* interchangeably as a notion of heterogeneity. We design experiments on three types of multi-modal data: (i) image/label pair— **FashionMNIST** images and labels; (ii) image/image pair—synthesized bimodal MNIST+MNIST datasets by pairing two different digits from **MNIST** as {(0,9),(1,8),(2,7),(3,6),(4,5)}; (iii) standard multi-modal dataset **CMU-MOSI** (Zadeh et al., 2018) including visual, textual and acoustic signals. See Appendix E for more results. For all we use the standard training/validation/test split and all masking follows MCAR. We evaluate the performance on labels with **PFC** (proportion of falsely classified attributes), images with **MSE** (mean-squared error) averaged over pixels and other attributes with **MSE**.

Data Imputation. Table 3 demonstrates VSAE can achieve superior performance for multi-modal data imputation on all modalities with lower variance. Fig. 3 presents the qualitative results of imputations on MNIST+MNIST image pairs. To demonstrate robustness to missing ratio, we conducted experiments with missing ratio of 0.3, 0.5, 0.7 on the MNIST+MNIST dataset and the sum errors by VSAE are 0.1371±0.0001, 0.1376±0.0002 and 0.1379±0.0001 respectively. This indicates that VSAE also stays robust under different missing ratios for multi-modal datasets.

Data Generation. Fig. 4 shows that VSAE is capable of generating image-based attributes following the underlying correlation. The learning process does not require any supervision and can be effectively carried out with only access to partially-observed data. We believe the underlying mechanism of selective proposal distribution benefits the performance. Though unobserved attributes in one data sample are not available even during training, they could be the observed attributes in others. Thus, the collective proposal networks are able to construct a stochastic mapping from observable to unobservable information among the whole training set. The separate structure of attributive and collective proposal network enforces VSAE to attend to the observed attributes, by ignoring unobserved attributes and performing efficient heterogeneity aggregation in the latent space. Thus it shows consistent robustness to various missing ratios. In contrast, baselines primarily approximate the posterior by a single proposal distribution inferred straight from the whole input. As a result baselines rely heavily on neural networks to extract expressive information from the data, which can be dominated by missing information.

5.3 Ablation Study

In Table 4, we perform extensive ablation studies to demonstrate the effectiveness of three critical modules: selective proposal network, mask distribution modeling, and expectation over generated $\mathbf{x_u}$ during training.

Selective proposal distribution. We compare VSAE to its variants with only collective or only attributive proposal network, which shows both are essential to the improvement. Empirically we observe the collective structure tends to outperform its attributive counterpart on *Yeast* but reversely worse on *Whitewine*. In fact, if we analyze the datasets, *Yeast* has highly correlated attributes (e.g. scores from different methods for the same task), whereas *Whitewine* has more

independent attributes (e.g. density, acidity or sugar).

Modeling the mask. We exclude mask conditional log-likelihood and make data decoder conditioned on the input mask. We observe the performance drops for non-MCAR, and do not observe significant improvement for MCAR. It indicates the mask reconstruction does not sacrifice model capacity in the case of MCAR, where the mask distribution can be easily learned.

Expectation over generated $\mathbf{x_u}$. This learning procedure mitigates the intractability of $\mathbf{x_u}$, therefore allowing to model the dependence between $\mathbf{x_u}$ and $\mathbf{m}$. We first consider removing this expectation procedure and ignoring $\mathbf{x_u}$ during training. The result (No EM) indicates this procedure plays an important role in NMAR ($\mathbf{x_u}$ depends on $\mathbf{m}$) and does not significantly sacrifice the performance of MCAR and MAR. Compared to larger sampling size (of 1000), the result shows our proposed method (which generates 100 samples of $\mathbf{x_u}$ to take expectation during training) is sufficient to obtain a good estimation of the expectation.

6 Conclusion

In this work, we propose VSAE, a novel latent variable model to learn from partially-observed heterogeneous data. The proposed VSAE handles missingness effectively by introducing a selective proposal distribution which is factorized w.r.t the data attributes. Further, VSAE is a general framework and capable of performing multiple tasks including data imputation, data generation and mask generation. Extensive experiments with comparison to state-of-the-art deep latent variable models demonstrated the effectiveness of VSAE on a variety of tasks. We summarize our contributions in the partially-observed heterogeneous setting as follows:

Heterogeneity. The factorization w.r.t. attributes in the latent space reduces the negative impact from the heterogeneity in the raw data space.

Partial observations. VSAE approximates the true posterior given partially-observed data with a novel selective proposal distribution. The automatic encoder selection between observed and unobserved attributes enables VSAE to ignore noisy information and learn from partial observations.

No MCAR assumption. The independence assumption between data and mask can be restrictive. VSAE relaxes this assumption and models the joint distribution of data and mask together. To handle the intractablility of $\mathbf{x_u}$, an expected lower bound is maximized during training.

References

Melissa J. Azur, Elizabeth A. Stuart, Constantine Frangakis, and Philip J. Leaf. Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, 2011.

Philip Bachman and Doina Precup. Data generation as sequential decision making. In *Advances in Neural Information Processing Systems*, 2015.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *arXiv*, 2017.

Yoshiaki Bando, Masato Mimura, Katsutoshi Itoyama, Kazuyoshi Yoshii, and Tatsuya Kawahara. Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization. In *International Conference on Acoustics, Speech and Signal Processing*, 2018.

Stef Buuren and Karin Groothuis-Oudshoorn. MICE: multivariate imputation by chained equations in R. *Journal of statistical software*, 2010.

Lluís Castrejón, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Learning aligned cross-modal representations from weakly aligned data. In *Conference on Computer Vision and Pattern Recognition*, 2016.

Mark Collier, Alfredo Nazabal, and Christopher K. I. Williams. VAEs in the presence of missing data. *arXiv*, 2020.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 1977.

Zoubin Ghahramani and Michael I. Jordan. Supervised learning from incomplete data via an EM approach. In *Advances in Neural Information Processing Systems*. 1994.

Lovedeep Gondara and Ke Wang. MIDA: Multiple imputation using denoising autoencoders. *arXiv*, 2018.

Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein GANs. In *Advances in Neural Information Processing Systems*, 2017.

Jingrui He. Learning from data heterogeneity: Algorithms and applications. In *International Joint Conference on Artificial Intelligence*, 2017.

Oleg Ivanov, Michael Figurnov, and Dmitry P. Vetrov. Variational autoencoder with arbitrary conditioning. In *International Conference on Learning Representations*, 2019.

Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning*, 2017.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Steven Cheng-Xian Li, Bo Jiang, and Benjamin M. Marlin. MisGAN: Learning from incomplete data with generative adversarial networks. In *International Conference on Learning Representations*, 2019.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2019.

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, José Miguel Hernández-Lobato, Sebastian Nowozin, and Cheng Zhang. EDDI: efficient dynamic discovery of high-value information with Partial VAE. In *International Conference on Machine Learning*, 2019.

Pierre-Alexandre Mattei and Jes Frellsen. MIWAE: deep generative modelling and imputation of incomplete data sets. In *International Conference on Machine Learning*, 2019.

Lars M. Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *International Conference on Machine Learning*, 2017.

Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *Conference on Computer Vision and Pattern Recognition*, 2016.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *International Conference on Multimodal Interfaces*, 2011.

Alfredo Nazabal, Pablo M. Olmos, Zoubin Ghahramani, and Isabel Valera. Handling incomplete heterogeneous data using vaes. *arXiv*, 2020.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, 2014.

Joseph L. Schafer. *Analysis of Incomplete Multivariate Data*. London, 1997.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, 2015.

Daniel J. Stekhoven and Peter Bühlmann. MissForest: non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 2011.

Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint multimodal learning with deep generative models. *arXiv*, 2016.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. In *International Conference on Learning Representations*, 2019.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning*, 2008.

Mike Wu and Noah D. Goodman. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems*, 2018.

Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick. Improved variational autoencoders for text modeling using dilated convolutions. In *International Conference on Machine Learning*, 2017.

Joonyoung Yi, Juhyuk Lee, Sung Ju Hwang, and Eunho Yang. Sparsity normalization: Stabilizing the expected outputs of deep networks. *arXiv*, 2019.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GAIN: missing data imputation using generative adversarial nets. In *International Conference on Machine Learning*, 2018.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. In *AAAI Conference on Artificial Intelligence*, 2018.

Amir Zadeh, Yao-Chong Lim, Paul Pu Liang, and Louis-Philippe Morency. Variational Auto-Decoder: Neural generative modeling from partial data. *arXiv*, 2019.

Yu Zhang, Ying Wei, and Qiang Yang. Learning to multitask. In *Advances in Neural Information Processing Systems*, 2018.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *International Conference on Computer Vision*, 2017.