Seminar Thesis

Learning Disentangled Representations with Semi-Supervised Deep Generative Models

Department of Statistics
Ludwig-Maximilians-Universität München

David B. Hoffmann

Munich, February 6th, 2026

Supervised by M. Sc. Simon Rittel

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Contents

1 Introduction

2   Background and Formulation

2.1   Disentanglement in Variational Auto-Encoders

Variational Auto-Encoders (VAEs) are generative models that learn a deep latent variable model by maximizing a lower bound on the marginal likelihood of the data (**??**). Given a dataset $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$, we assume the data is generated by a random process involving an unobserved continuous random variable z. The generative process consists of a prior distribution $p_\theta(z)$ and a conditional likelihood $p_\theta(x|z)$, typically parametrised by a neural network (the decoder) with parameters θ.

Since the true posterior $p_\theta(z|x)$ is generally intractable, VAEs introduce an approximate posterior $q_\phi(z|x)$, parametrised by a separate neural network (the encoder) with parameters ϕ. The model is trained by maximizing the evidence lower bound (ELBO) on the marginal log-likelihood $\log p_\theta(x)$:

$$\log p_\theta(x) \geq \mathcal{L}_{\text{ELBO}}(x; \theta, \phi) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)\|p(z)) \tag{1}$$

The first term is the reconstruction error, encouraging the decoder to recover the data from the latent code. The second term is the Kullback-Leibler (KL) divergence, which regularizes the approximate posterior to be close to the prior, typically assumed to be a standard multivariate Gaussian $p(z) = \mathcal{N}(0, I)$.

To allow for backpropagation through the stochastic sampling process, VAEs utilize the reparametrisation trick. Instead of sampling z directly from $q_\phi(z|x) = \mathcal{N}(\mu, \sigma^2)$, we sample an auxiliary noise variable $\epsilon \sim \mathcal{N}(0, I)$ and compute:

$$z = \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon \tag{2}$$

where $\odot$ denotes the element-wise product.

A key goal in representation learning is disentanglement, where individual latent units are sensitive to changes in single generative factors of the data while being invariant to others (**?**). Formally, a representation is disentangled if it factorizes into independent subspaces corresponding to underlying factors of variation.

Standard VAEs often fail to learn disentangled representations due to the lack of explicit constraints on the latent structure. To address this, the β-VAE framework **?** introduces a hyperparameter β to weight the KL divergence term:

$$\mathcal{L}_{\beta\text{-VAE}} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \beta D_{KL}(q_\phi(z|x)\|p(z)) \tag{3}$$

Setting $\beta > 1$ imposes a stronger constraint on the latent bottleneck, encouraging the learned distribution $q_\phi(z|x)$ to align with the isotropic unit Gaussian prior. Since the prior has independent components, this pressure encourages the latent dimensions to become statistically independent, thereby promoting disentanglement, though often at the cost of reconstruction quality.

2.2   Degrees of Supervision

While standard supervised learning relies on a fully labelled dataset $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$, obtaining ground truth labels y is often expensive. Semi-supervised learning addresses

this by leveraging a small set of labelled data $\mathcal{D}_S$ alongside a much larger set of unlabelled data $\mathcal{D}_U = \{x^{(j)}\}_{j=1}^M$. The objective is to utilize the structural information inherent in $p(x)$ from $\mathcal{D}_U$ to improve the estimation of the conditional distribution $p(y|x)$, typically under the assumption that the decision boundary lies in low-density regions of the input space.

In contrast, weak supervision relaxes the requirement for exact ground truth labels, utilizing instead a dataset $\mathcal{D}_W = \{(x^{(i)}, \tilde{y}^{(i)})\}_{i=1}^N$ where $\tilde{y}$ represents a noisy, limited, or heuristic approximation of the true label y. A prominent subset is pair learning, where supervision is provided as pairwise constraints $\mathcal{D}_P = \{(x^{(i)}, x^{(j)}, r^{(ij)})\}$. Here, $r^{(ij)}$ denotes a binary relationship (e.g., must-link or cannot-link) between samples, guiding the model to learn a representation that respects these relative similarities rather than absolute class assignments.

2.3 Semi-Supervised Variational Auto-Encoders

To leverage the generalization benefits of generative modelling for classification tasks, ? extend the VAE framework to the semi-supervised setting. They propose three approaches: the *M1* model, which separates feature learning from downstream classification; the *M2* model, which integrates the class label directly into the latent generative process; and a stacked *M1+M2* architecture. We focus here on the *M2* model, where the data x is generated by both a latent class variable y and a continuous latent variable z.

The model describes the data x as being generated by a latent class variable y and a continuous latent variable z. The generative process factorizes as:

$$p_\theta(x, y, z) = p_\theta(x|y, z)p(y)p(z) \tag{4}$$

where $p(y) = \text{Cat}(y|\pi)$ and $p(z) = \mathcal{N}(z|0, I)$. In this prior specification, y and z are marginally independent. The likelihood $p_\theta(x|y, z)$ is parameterized by a deep neural network that takes both y and z as inputs to generate the observation x.

Crucially, the inference structure introduced to approximate the intractable posterior $p(z, y|x)$ utilizes a specific factorization that introduces a dependency between the latent variables. The recognition model is specified as:

$$q_\phi(z, y|x) = q_\phi(z|x, y)q_\phi(y|x) \tag{5}$$

Here, $q_\phi(y|x)$ is modelled as a categorical distribution, which allows it to be used as a classifier, while $q_\phi(z|x, y)$ is a Gaussian distribution where the parameters (mean and variance) are functions of both the data x and the label y. This dependency structure in the inference network allows the model to learn a class-conditional latent distribution for z, enabling it to effectively separate style z from class y.

In the inference process, however, these variables are coupled. The approximate posterior for the continuous variable is defined as $q_\phi(z|x, y)$. By conditioning on y, the network encourages z to capture residual variations orthogonal to class identity.

For the supervised dataset $\mathcal{D}_S = \{(x^m, y^m)\}$, both x and y are observed. The supervised objective $\mathcal{L}_S(\theta, \phi; x, y)$ consists of two main components: a generative term which is the ELBO on the joint distribution of x, y, and z as well as a discriminative term weighted by α. By expanding the generative term using the factorization

$p_\theta(x, y, z) = p_\theta(x|y, z)p(y)p(z)$, we obtain a reconstruction term which encourages the decoder output to closely match the encoder input, a KL regularization that encourages the style latent z to follow the prior, and the label prior $\log p(y)$:

$$\mathcal{L}_S(\theta, \phi; x, y) = \underbrace{\mathbb{E}_{q_\phi(z|x,y)}\left[\log \frac{p_\theta(x, y, z)}{q_\phi(z|x, y)}\right]}_{\text{Generative (ELBO on joint } x,y)} + \underbrace{\alpha \log q_\phi(y|x)}_{\text{Discriminative}} \tag{6}$$

$$= \mathbb{E}_{q_\phi(z|x,y)}\bigg[\log p_\theta(x|y, z) + \log p(y)$$

$$+ \log p(z) - \log q_\phi(z|x, y)\bigg] + \alpha \log q_\phi(y|x) \tag{7}$$

$$= \underbrace{\mathbb{E}_{q_\phi(z|x,y)}[\log p_\theta(x|y, z)]}_{\text{Reconstruction}} - \underbrace{D_{KL}(q_\phi(z|x, y)\|p(z))}_{\text{Style Regularisation}}$$

$$+ \underbrace{\log p(y)}_{\text{Label Prior}} + \underbrace{\alpha \log q_\phi(y|x)}_{\text{Discriminative}} \tag{8}$$

For the unsupervised dataset $\mathcal{D}_U = \{x^n\}$, the label y is treated as a latent variable. We marginalize over all classes to derive the unsupervised bound $\mathcal{L}_U(\theta, \phi; x)$:

$$\mathcal{L}_U(\theta, \phi; x) = \sum_y q_\phi(y|x)\left(\mathbb{E}_{q_\phi(z|x,y)}\left[\log \frac{p_\theta(x, y, z)}{q_\phi(z|x, y)}\right]\right) + \mathcal{H}(q_\phi(y|x)) \tag{9}$$

where $q_\phi(y|x)$ is the recognition network acting as a discriminative classifier.

The final training objective combines these bounds over both datasets:

$$\mathcal{L}(\theta, \phi; \mathcal{D}_U, \mathcal{D}_S) = \sum_{x \in \mathcal{D}_U} \mathcal{L}_U(\theta, \phi; x) + \gamma \sum_{(x,y) \in \mathcal{D}_S} \mathcal{L}_S(\theta, \phi; x, y) \tag{10}$$

Where γ is used to balance the contribution of the supervised data. Note that this additional weighting parameter was not part of the original formulation from **?** which solely relied on α.

2.4 Importance Sampling

Importance sampling is a Monte Carlo method used to estimate properties of a target distribution $p(x)$ that is difficult to sample from directly, by utilizing a simpler proposal distribution $q(x)$.

Given a target density $p(x)$ and a function $f(x)$, we wish to estimate the expectation $\mathbb{E}_p[f(x)]$. We introduce a proposal distribution $q(x)$ such that $q(x) > 0$ whenever $p(x)f(x) \neq 0$ (i.e., the support of q covers the support of pf). By multiplying and dividing by $q(x)$, we rewrite the expectation as:

$$\mathbb{E}_{x \sim p}[f(x)] = \int f(x)p(x)\, dx = \int f(x)\frac{p(x)}{q(x)}q(x)\, dx = \mathbb{E}_{x \sim q}\left[f(x)\frac{p(x)}{q(x)}\right] \tag{11}$$

The term $w(x) = \frac{p(x)}{q(x)}$ is known as the importance weight. In practice, we draw K samples $\{x^{(k)}\}_{k=1}^{K}$ from the proposal distribution $q(x)$ and approximate the expectation using the sample mean:

$$\mathbb{E}_p[f(x)] \approx \frac{1}{K} \sum_{k=1}^{K} f(x^{(k)}) w(x^{(k)}) \tag{12}$$

This standard estimator is unbiased, provided $p(x)$ is a normalized density.

However, in many applications, the target distribution is only known up to an intractable normalizing constant Z, i.e., $p(x) = \frac{\tilde{p}(x)}{Z}$. In this scenario, self-normalized importance sampling is used. We calculate unnormalized weights $\tilde{w}(x) = \frac{\tilde{p}(x)}{q(x)}$ and approximate the expectation using the ratio of estimators:

$$\mathbb{E}_p[f(x)] \approx \frac{\sum_{k=1}^{K} \tilde{w}(x^{(k)}) f(x^{(k)})}{\sum_{k=1}^{K} \tilde{w}(x^{(k)})} \tag{13}$$

Unlike the standard estimator, the self-normalized estimator is biased for finite K (with bias of order $1/K$), but consistent, converging to the true expectation as $K \to \infty$.

3 Semi-Supervised Disentanglement with Arbitrary Dependencies

While the *M2* model by **?** provides a powerful framework for semi-supervised learning, it imposes a specific graphical structure: the latent variables y (label) and z (style) are assumed to be independent in the prior, and the recognition model is required to factorise as $q_\phi(y, z|x) = q_\phi(y|x)q_\phi(z|x, y)$. **?** generalise this approach to a broader class of partially-specified graphical models.

In this framework, the modeller specifies a graphical model where a subset of variables y are semantically meaningful (and partially observed), while the remaining variables z are left unstructured to capture nuisance variations. Unlike the *M2* model, this framework allows for arbitrary dependency structures in both the generative model $p_\theta(x, y, z)$ and the approximate posterior $q_\phi(y, z|x)$. This flexibility enables the modelling of complex relationships, such as hierarchical dependencies or causal structures between labels and style variables, which are often necessary for disentanglement in real world datasets.

3.1 The Generalised Variational Objective

The core challenge in training these generalised models lies in the supervised objective as defined in Equation 6. For the unsupervised term $\mathcal{L}_U$ (Equation 9), standard variational inference applies. However, for the supervised term $\mathcal{L}_S$, a difficulty arises when the recognition model $q_\phi(y, z|x)$ does not factorise conveniently. Specifically, evaluating the supervised lower bound requires sampling from the conditional posterior $q_\phi(z|x, y)$. For arbitrary graphical structures, this conditional distribution may not be available in closed form or easy to sample from directly.

To address this, **?** propose an importance sampling estimator that alleviates the need to sample from $q_\phi(z|x, y)$ directly. Instead, samples are drawn from an unconditioned proposal distribution, the marginal encoder $q_\phi(z|x))$, and reweighted.

They start by rewriting the supervised objective using the fact that $q_\phi(z|x, y)$ factorizes to $\frac{q_\phi(y, z|x)}{q_\phi(y|x)}$ as follows

$$\mathcal{L}_S(\theta, \phi; x, y) = \alpha \log q_\phi(y|x) + \mathbb{E}_{q_\phi(z|x,y)}\left[\log \frac{p_\theta(x, y, z)}{q_\phi(z|x, y)}\right] \tag{14}$$

$$= \alpha \log q_\phi(y|x) + \mathbb{E}_{q_\phi(z|x,y)}\left[\log \frac{p_\theta(x, y, z)}{q_\phi(y, z|x)} + \log q_\phi(y|x)\right] \tag{15}$$

$$= (1 + \alpha) \log q_\phi(y|x) + \mathbb{E}_{q_\phi(z|x,y)}\left[\log \frac{p_\theta(x, y, z)}{q_\phi(y, z|x)}\right] \tag{16}$$

This formulation removes the need to evaluate $q_\phi(z|x, y)$ directly. To approximate the expectation, they employ self-normalised importance sampling as introduced in subsection 2.4. Instead of sampling from the conditional $q_\phi(z|x, y)$, they sample proposals z^s from the unconditioned encoder distribution $q_\phi(z|x)$. The unnormalised importance weights w^s are defined as the ratio of the target joint distribution to the proposal

$$w^s = \frac{q_\phi(y, z^s|x)}{q_\phi(z^s|x)} \tag{17}$$

where $z^s \sim q_\phi(z|x)$. Using these weights, the expectation is approximated as

$$\mathbb{E}_{q_\phi(z|x,y)}\left[\log \frac{p_\theta(x, y, z)}{q_\phi(y, z|x)}\right] \approx \frac{1}{S}\sum_{s=1}^{S}\frac{w^s}{Z}\log \frac{p_\theta(x, y, z^s)}{q_\phi(y, z^s|x)} \tag{18}$$

where $Z = \frac{1}{S}\sum_{s=1}^{S}w^s$ is the normalisation constant.
Furthermore, the term $\log q_\phi(y|x)$ is itself approximated using a Monte Carlo estimator of the lower bound normally used in maximum likelihood estimation

$$\log q_\phi(y|x) \geq \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{q_\phi(y, z|x)}{q_\phi(z|x)}\right] \approx \frac{1}{S}\sum_{s=1}^{S}\log w^s \tag{19}$$

By combining the importance sampling estimator for the expectation (Equation 18) and the lower bound estimator for the log-likelihood (Equation 19), they obtain the final estimator for the supervised objective

$$\hat{\mathcal{L}}_S(\theta, \phi; x, y) := \frac{1}{S}\sum_{s=1}^{S}\frac{w^s}{Z}\log \frac{p_\theta(x, y, z^s)}{q_\phi(y, z^s|x)} + (1 + \alpha)\log w^s \tag{20}$$

This formulation ensures that the discriminative power of the model is maximised using the same samples z^s and weights w^s for both the generative and discriminative components.

3.2 Relation to Kingma's M2 Model

The framework by **?** generalizes the *M2* model. We can recover the exact *M2* objective by restricting the dependency structure. If we enforce the factorisation:

$$q_\phi(y, z|x) = q_\phi(y|x)q_\phi(z|x, y) \tag{21}$$

then the importance weights simplify to constants with respect to z

$$w^s = \frac{q_\phi(y|x)q_\phi(z^s|x, y)}{q_\phi(z^s|x, y)} = q_\phi(y|x) \tag{22}$$

Substituting this back into the estimator yields the standard *M2* objective. However, the generalised importance sampling formulation allows for alternative factorisations, such as $q_\phi(y, z|x) = q_\phi(y|x, z)q_\phi(z|x)$, where the label is predicted from the latent style z. This inverted structure effectively treats the latent variable z as a sufficient statistic for the label, enforcing a stronger form of disentanglement where z must contain all information necessary to predict y.

3.3 Graphical Model and Stochastic Implementation

To perform gradient ascent, the generative and recognition models are mapped onto a stochastic computation graph where each node forms a sub-graph. The generative model assumes a factorization $p_\theta(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z})p(\mathbf{y})p(\mathbf{z})$, while the recognition model utilizes a conditional dependency structure $q_\phi(\mathbf{y}, \mathbf{z}|\mathbf{x}) = q_{\phi_z}(\mathbf{z}|\mathbf{y}, \mathbf{x})q_{\phi_y}(\mathbf{y}|\mathbf{x})$ to disentangle the digit label from the handwriting style.

The implementation handles supervision through three specific node types. First, for the fully supervised variable $\mathbf{x}$, they compute the likelihood $p_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z}) = \mathcal{N}(\mathbf{x}; \eta_\theta(\mathbf{y}, \mathbf{z}))$, where η_θ is a neural network returning the distribution parameters. Second, the unobserved latent variable $\mathbf{z}$ requires computing both the prior $p(\mathbf{z})$ and the conditional $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y})$. This utilizes the Gaussian reparametrisation trick $\mathbf{z} = g(\epsilon, \lambda_{\phi_z}(\mathbf{x}, \mathbf{y}))$ where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to ensure differentiability. Finally, the partially observed variable $\mathbf{y}$ is treated as observed when labels are available. When unavailable, it is sampled from $q_{\phi_y}(\mathbf{y}|\mathbf{x})$ using a Gumbel-Softmax relaxation to allow for gradient estimation on the discrete distribution.
? implement these different node types in a stochastic machine learning framework called ProbTorch which is available at: `https://github.com/probtorch/probtorch`.

3.4 Experiments and Findings

? evaluate the proposed framework across three distinct domains to demonstrate its ability to learn disentangled representations under partial supervision and varying structural complexity.

The authors do not provide full details about the implementation of their experiments. However, they do state that they use a linear layers for Modified National Institute of Standards and Technology (MNIST), a convolutional network for Street View House Numbers (SVHN) and the Yale-B datasets and a gated recurrent unit for the Multi-MNIST experiment. For the training process they use Adam with the "default learning rate and momentum correction terms". **?** vaguely state that depending on the dataset a bach size between 100-700 is used. Experiment specific information is mentioned in the following subsections.

3.4.1 Benchmark Classification: MNIST and SVHN

To validate the generalised objective against the standard *M2* model (**?**), the authors first applied the framework to the MNIST (**?**) and SVHN (**?**) datasets. The graphical model structure mirrored the *M2* setup, where the latent space consists of a partially observed class label y and an unobserved style variable z. They state that they use $\gamma = 1$ and introduce a supervision rate $\rho = \frac{\gamma M}{N + \gamma M}$. The α parameter is not specified.

On the MNIST dataset classification performance is evaluated for supervised set sizes of 100, 600, 1000 and 3000 our of the 50.000 unsupervised samples, where they marginally outperform the baseline *M2* model. In the more complex SVHN domain the authors compare classification results for 1000 and 3000 labels out of the 70,000 unsupervised samples. They find that their Semi-Supervised VAE (SSVAE) performs similar compared to the two-stage *M1+M2* approach despite using a single-stage training process (**?**). Qualitatively, the results demonstrated a clean separation of style and content. By fixing the style

latent z and varying the label y, the model generated analogies where the same handwriting style was consistently applied to different digits. Conversely, fixing y and varying z produced variations in stroke width and slant while maintaining the digit identity. Furthermore, the authors analysed the effect of the supervision weight γ and found that for sparsely labelled datasets ($M \ll N$), over-weighting the supervised term ($\gamma > 1$) significantly improved generalisation on the test set, although excessive weighting eventually led to overfitting. Here they offer conflicting statements about α where the figure description states that $\alpha = 50$ for MNIST and $\alpha = 70$ for SVHN while the discussion of the results of the supervision rate mentions that they used $\alpha = 0.1/\rho$ in line with **?**. For a more detailed description of experimental results refer to Figure 3 of **?**.

3.4.2   Intrinsic Faces: Disentangling Continuous Factors

Moving beyond categorical labels, the authors utilised the Yale-B dataset (**?**) to learn disentangled representations of *Identity* (categorical) and *Lighting* (continuous). The generative model defined latent variables for identity, lighting, shading, and reflectance, with partial supervision provided only for identity and lighting direction. The authors use a supervision rate of $\rho = 0.5$ with 6 supervised and 32 unsupervised data points corresponding which implies $\gamma \approx 5.33$. α is set to 10. The authors demonstrate that the model disentangled identity from lighting conditions with qualitative samples where they manipulate lighting direction on a specific identity's face without altering facial features. Quantitatively, the framework outperformed an existing baseline on the identity classification task, achieving a semi-supervised error rate of 3.5% compared to approximately 30% for the baseline method by **?**. For the regression of the continuous lighting parameter, the semi-supervised model achieved an angular error of 17.6%, compared to approximately 10% for the fully supervised baseline from **?**, demonstrating the efficacy of the objective for continuous latents even with limited supervision. For a more detailed description of experimental results refer to Figure 4 of **?**.

3.4.3   Multi-MNIST: Stochastic Dimensionality and Compositionality

The final experiment addressed a complex counting task using a custom Multi-MNIST dataset, where each image contained a varying number (1–3) of MNIST digits. This required a model with stochastic dimensionality, where the number of latent variables is itself a random variable K. The authors proposed a recursive generative model that sequentially samples digits and places them on a canvas using Spatial Transformer Networks. Fir this experiment neither α, γ or ρ are reported by the authors. The model learned to reliably predict the number of digits K and successfully decomposed overlapping digits into their constituent parts and background. Additionally, the authors compared a flat model against a nested model that incorporated a pre-trained MNIST auto-encoder as a sub-component. This nested configuration demonstrated the framework's capacity for modular model design and transfer learning. While the flat model achieved higher raw counting accuracy, the nested model still effectively leveraged the pre-learned constituent representations to reconstruct complex multi-digit scenes and decompose inputs into coherent individual digits. For a more detailed description of experimental results refer to Figure 6 of **?**.

4 Extensions and Limitations

Several works build on the results form **?** and extend it to specific domains or in a general context. However, there are also some limitations and critiques in the subsequent literature. Both aspects are presented in the following sections. Relevant sources where systematically identified as outlined in Appendix A.

4.1 Own Criticism

In the following we present several issues in presentation **?** as well as unverified claims and minor critiques.

The presentation of the SSVAE framework in **?**, while correct, at points leaves the reader guessing at specific details, disrupting the reading flow. As such, some figures do not have captions and are never referenced. Other figures are referenced without explaining their interpretation such as the meaning of shading in the nodes of the computational graph in Figure 1 and 5 of **?**. It is never explained that the shading corresponds to the degree to which the variables are supervised. Furthermore, some variables are never defined such as the n in the relationship

$$(n \cdot l) \times r + \epsilon \tag{23}$$

of the supervised variables in the Yale-B dataset (**?**) experiment we explain in subsubsection 3.4.2. The meaning of this variable only becomes clear after reading the cited work of **?** which declares it as the normal map.

Further, **?** claim a number of properties and findings that are not shown theoretically or empirically. One example of this is their in the introduction that "a representation that has some factorisable structure, and consistent semantics associated to different parts, is more likely to generalise to a new task". This is not verified by their experiment or any cited source. In a later paper **?** investigate this claim and find no evidence to support it as further explained in subsubsection 4.2.1.

The authors introduce SSVAE as a framework for learning disentangled representations. **?** claim that it provides the user with the ability to precisely specify axes of variation and their dependencies. Specifically, in their Yale-B dataset experiment they claim to learn the relationship between lighting, shading and reflectance from Equation 23. While they qualitatively demonstrate disentangling as well as classification and regression performance, the learning of this relationship is not shown.

While the SSVAE framework generalizes to arbitrary dependency structures, its practical application in real-world settings with multiple partially observed variables reveals significant limitations. The current formulation relies on a binary distinction between unsupervised data $\mathcal{D}_U$ and supervised data $\mathcal{D}_S$, assuming a consistent set of available labels for the supervised portion. However, in complex real-world datasets involving K distinct supervised variables $\mathbf{y} = \{y_1, \ldots, y_K\}$, supervision is often heterogeneous; a datapoint $x^{(n)}$ may possess annotations for a random subset of variables $O^{(n)} \subseteq \mathbf{y}$, while remaining variables $U^{(n)} = \mathbf{y} \setminus O^{(n)}$ are missing.

Under the proposed framework, variables are treated as observed when available and sampled otherwise. Yet, implementing this dynamically in a batched stochastic gradient descent setting becomes non-trivial. The standard objective presented in Equation 10

breaks down because there is no single $\mathcal{L}_S$. Instead, the objective function effectively fractures into 2^K potential observation patterns. To handle this, the practitioner must manually specify unique loss components or intricate masking logic for each missingness pattern to correctly toggle between computing the likelihood $p(y_k)$ (for $y_k \in O^{(n)}$) and performing importance sampling using $q_\phi(y_k|x)$ (for $y_k \in U^{(n)}$). This combinatorial explosion necessitates a bespoke and brittle implementation of the stochastic computation graph, undermining the flexibility intended by the general framework.

Additionally, while it is considered a good practice to reproduce experiments for methods presented in other papers instead of directly copying their results. This is not done by **?**. For their MNIST/SVHN and Yale-B experiments (outlined in subsubsection 3.4.1 and subsubsection 3.4.2 respectively) they merely copy experimental result values from **?** and **?** instead of confirming the performance in their own experiments. Moreover, the definition of hyperparameters such as α and γ is not always clear as alluded to in subsection 3.4. Another minor error is the wrong column description in Figure 6 of **?** which states that the table on the bottom right reports count error, while the table actually contains accuracy values as correctly pointed out in the corresponding caption.

Lastly, in the same experiments, **?** do not mention which exact values they report. They state that they report error rates but not which metric they use. This complicates comparability to other methods and isolated interpretation of their results.

4.2 General Critiques

In the following the semi-supervised disentanglement approach from **?** is confirmed by results from **?**, which demonstrate the impossibility of stable unsupervised disentanglement. **?** offer a general critique of the independent and identically distributed (i.i.d.) assumption in VAEs directly mentioning **?**.

4.2.1 Impossibility of Unsupervised Disentanglement

In their paper, "Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations", **?** presented a proof that unsupervised disentanglement is fundamentally impossible without inductive biases on both the model and the data.

The proof relies on the geometry of the latent space and the properties of the Gaussian prior used in VAEs. Assuming a ground-truth generative process $x = f(z)$ where $z \sim P(z)$, then if $P(z)$ is a standard multivariate Gaussian, it is rotationally symmetric. That is, for any orthogonal rotation matrix R, the distribution of Rz is identical to the distribution of z.

Given a new generative function $f'(z) = f(R^T z)$. The marginal distribution of observations $p(x)$ produced by this new model is identical to the original:

$$ p(x) = \int p(x|z)p(z)dz = \int p(x|R^T z)p(z)dz $$

Since the unsupervised ELBO only depends on the marginal likelihood of the data $p(x)$, it cannot distinguish between the true disentangled model $f(z)$ and the entangled model $f'(z)$.

In their experiments **?** "do not find any evidence that the considered models can be used to reliably learn disentangled representations in an unsupervised manner as random seeds and hyperparameters seem to matter more than the model choice. Furthermore, good trained models seemingly cannot be identified without access to ground-truth labels even if we are allowed to transfer good hyperparameter values across data sets". Furthermore, they state that for the considered models and data sets, "increased disentanglement does not seem to lead to a decreased sample complexity of learning for downstream tasks".

This result confirms the necessity of some form of inductive bias for stable latent representations as proposed by **?**. **?** anchor the variable y with labels, however, the variable z is essentially learned in an unsupervised manner relative to the style factors. **?**'s proof implies that there is no guarantee that z will factorize into meaningful sub-components purely based on the VAE objective. The model is free to learn any rotation of these style factors that explains the data.

? further highlight that they couldn't find any evidence that disentangled representation lead to better downstream performance, shifting the focus more towards interpretability and fairness. While **?** claim the contrary, namely that consistent and factorisable representation are more likely to generalise to a new task, they never formally prove or demonstrate this empirically in their paper.

4.2.2 The Limitation of Isotropic Priors

? critique the use of overly simplistic i.i.d. Gaussian priors. While **?** move away from such priors by composing latent graphical models with deep likelihoods, **?** argue that the framework still often relies on conditional independence assumptions that may induce excessive regularization. **?** point out that for many important datasets—such as time-series of images, medical scans of the same patient, or rotated views of an object—the samples exhibit structured correlations that are better captured through the prior's covariance.

Using a prior that ignores these sample-to-sample correlations is a model misspecification that forces the encoder to discard the correlation structure. **?** introduced the Gaussian Process Prior VAE (GPPVAE), which replaces the independent prior with a Gaussian Process:

$$\mathbf{z} \sim \mathcal{GP}(0, K(X, X'))$$

Here, the covariance kernel K explicitly models the correlation between samples (e.g., temporal proximity or identity). This allows the model to disentangle "object identity" from "view" by leveraging the kernel structure, a form of disentanglement that **?**'s graphical approach typically handles through semi-supervised labels and specific message-passing variational families.

? also address the limitations of the isotropic Gaussian prior, arguing that it fails to exert sufficient regularizing pressure to force effective disentanglement. To resolve the common problem of disentangling discrete factors of variation from continuous variables—a task that remains problematic for many contemporary approaches—they propose a structured decomposition of the ELBO objective. This modification allows for explicit control over the relative levels of disentanglement within different groups of latent variables. However, their analysis also highlights a significant optimization hurdle: discrete variables tend to exhibit higher likelihood values than continuous variables, which can introduce a bias that

skews the optimization process in semi-supervised generative models.

4.2.3 Unbounded Likelihoods and Mode Collapse

? investigated the maximum likelihood estimation for Deep Latent Variable Models (DLVMs), the class of models to which **?** semi-supervised VAE belongs. They proved that for continuous data, if the variance of the decoder's output distribution is learned without constraints, the likelihood function is unbounded.

Consider a decoder that outputs a Gaussian distribution $\mathcal{N}(\mu_\theta(z), \sigma_\theta^2(z))$. If the model can map a specific latent point z_i to exactly match a data point x_i (i.e., $\mu_\theta(z_i) \approx x_i$) and simultaneously drive the variance $\sigma_\theta^2(z_i)$ toward zero, the likelihood density $p(x_i \mid z_i)$ approaches infinity:

$$p(x_i|z_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \to \infty \text{ as } \sigma \to 0$$

This creates a "hole" in the optimization landscape where the model collapses to storing data points (overfitting) rather than learning a generalized generative process. This is a severe limitation for **?**s framework when applied to continuous data, as it can lead to numerical instability and meaningless latent representations where the "style" z degenerates into a lookup table index rather than a semantic factor. **?** suggest constraints on σ or specific architectural choices to ensure the existence of generalized maximum likelihood estimates.

4.3 Disentanglement Implications and Limitations

In the following, disentanglement and its limitations are presented.

4.3.1 Semantic Conflation Problem

?'s model feeds the label y and the latent z into the decoder: $p(x \mid y, z)$. The assumption is that y handles the class-specific information and z handles the rest. **?** argue that this rigid separation is flawed because labels often imply continuous characteristics. Specifically they state that "Originally motivated by the desiderata of semi–supervised classification, each label is given a corresponding latent variable of the same type (e.g. categorical), whose value is fixed to that of the label when the label is observed and imputed by the encoder when it is not. Though natural, we argue that this assumption is not just unnecessary but actively harmful from a representation-learning perspective, particularly in the context of performing manipulations. To allow manipulations, we want to learn latent factors that capture the characteristic information associated with a label, which is typically much richer than just the label value itself."

They provide an example pertaining to a dataset of face. Consider a face with the label "Young". **?** argue that age is not inherently discrete and that youth correlates with a number of continuous features such as smooth skin, specific facial structures, and hair density. In a standard VAE, the decoder sees $y = 1$ ("Young") and learns to generate these features. Consequently, the latent z is absolved of the responsibility to encode the degree of youth or the specific way youth manifests in that image. This is semantic

conflation: the discrete label "steals" the semantic content from the continuous latent space.

They state that this leads to a failure in disentanglement: if we want to manipulate the "age" of a face continuously, we cannot do so easily because the age information is locked inside the discrete variable y. Conversely, if we change the label y from "Young" to "Old" while keeping z fixed, the style z (which might encode "smiling") might be ignored or misinterpreted by the decoder because, so **?** argue, the correlations between style and age are broken.

To resolve this, they proposed the Characteristic Capturing VAE (CCVAE) which builds on the SSVAE proposed by **?**. They radically altered the graphical model. Instead of conditioning the generation on the label, they condition the inference of specific latents on the label. Specifically, they split the latent space into: (1) characteristic latents (z_c) which capture information correlated with the label and (2) salient latents (z_s) which capture the residual information.

The generative model is $p(x \mid z_c, z_s)$. Note that y is not an input to the decoder. Instead, an auxiliary classifier ensures that z_c is predictive of y: $\mathbb{E}[\log p(y \mid z_c)]$. By forcing z_c to predict y but feeding z_c (not y) to the decoder, the model must encode the semantic content of the label into the continuous space z_c. This avoids conflation. The label y acts as a supervisor for the structure of the latent space rather than a crutch for the decoder. **?** demonstrated that this leads to superior performance in "attribute traversal." For example, dealing with the attribute "Young," the CCVAE could generate smooth transitions from young to old faces by interpolating in z_c, whereas the baseline models could only flip the binary switch y, resulting in discontinuous and often lower-quality transitions.

? underline this critique by and state that "it still remains unclear how the use of supervision impacts the disentanglement learning". Which in part is also due to a lack of relevant metrics (**?**), making it hard to quantify how well disentanglement works and how the semi-supervised framework of **?** impacts it.

4.3.2 Breaking the ELBO Bottleneck

? identified a specific optimization pathology in loss function of SSVAEs (**?**), termed the *ELBO Bottleneck* and proposed SmootH-ELBO Optimal InTerpolation VAE (SHOT-VAE) as an extension of SSVAE to mitigate this limitation.

In SSVAEs, the objective combines the ELBO and a classification loss. **?** observed that as training progresses, the ELBO term often plateaus before the inference accuracy is maximized. The standard ELBO does not strictly penalize misclassifications in the latent space as long as the reconstruction is good. This leads to a disconnect: the model might reconstruct the image of a "9" perfectly well even if the latent code z and label y are slightly misaligned or ambiguous.

In the SSVAEs framework proposed by **?**, the latent space is explicitly partitioned into a discrete label y and a continuous style variable z. While this separation is designed to "anchor" semantics, **?** identify a fundamental optimization pathology in this joint objective: the *ELBO Bottleneck*. Because the generative term $p(x|y, z)$ can achieve high likelihood by absorbing label-relevant information into the "unspecified" style variable z, the model often reaches an ELBO plateau where reconstruction is near-perfect despite poor classification accuracy. This indicates that the inductive bias intended by **?** is often

bypassed during training, as the standard ELBO lacks the gradient pressure to prevent z from "leaking" the information that should be exclusive to y.

? introduce SHOT-VAE with two key innovations (1) Smooth-ELBO which is an approximation that integrates the label predictive loss directly into the ELBO derivation, rather than treating it as an auxiliary loss. This aligns the generative and discriminative objectives more tightly. (2) Optimal Interpolation for which they utilized data augmentation in the latent space (mixup) to fill the gaps between class clusters. By enforcing linearity in the latent space (interpolating between a "1" and a "7" should yield a semantic blend), they break the ELBO bottleneck and force the encoder to learn a more robust, disentangled structure.

Empirically, SHOT-VAE achieved significant error rate reductions on CIFAR-10 (6.11% vs baseline) compared to standard SSVAEs from **?**, demonstrating that optimization dynamics are as critical as model architecture for disentanglement.

4.3.3 Differentiating Consistency and Restrictiveness

? introduce a refined definition of disentanglement that is made up of consistency and restrictiveness. Here consistency is the degree to which a representation is deterministic with respect to the ground-truth factors. If we fix the ground-truth factor (e.g., color), the latent code should ideally be constant. On the other hand, restrictiveness is the degree to which a single dimension of the representation encodes only one ground-truth factor. This prevents a single latent dimension from encoding both color and shape.

? state that **?** falsely claim that their method leads to disentangled results through semi-supervision. Instead is merely creates consistent representation which are not necessarily restrictive and hence not guaranteed to be disentangled. However, they relativise this by conceding that on real world data consistency and restrictiveness are often strongly correlated.

These definitions provide a finer granularity than the mutual information metrics used previously. Critically, observational metrics (like the Mutual Information Gap (MIG) (**?**) used in **?**) often fail to capture these causal properties accurately. The differentiation of disentanglement introduced by **?** paves the way for researchers to formally prove which types of weak supervision (e.g., restricted labeling, match-pairing) guarantee consistency or restrictiveness. This represents a theoretical maturation from **?**'s reliance on empirical validation via reconstruction.

4.4 Extensions in Supervision: From Semi to Weak

In response to the impossibility results from **?**, the research community pivoted. If pure unsupervised disentanglement is impossible, and full supervision is expensive, what is the minimal signal required? **?** proposed Weakly-Supervised Disentanglement as a robust extension to the semi-supervised paradigm. **?** further extend the concept of weakly-supervised VAEs to multimodal learning and inference.

4.4.1 The Weak Supervision Paradigm

While **?** highlight the need for an inductive bias to create stable representations. Several works (**????**) also critique the use of labels in the framework of **?**. These criticisms are summarized well by **?** who note that "semi-supervised approaches that require implicit or explicit knowledge about the true underlying factors of the data have excelled at disentangling" referring to **?**. However, they also point out that "ideally we would like to learn these in an unsupervised manner, due to the following reasons: 1. Humans are able to learn factors of variation unsupervised [**?**]. 2. Labels are costly as obtaining them requires a human in the loop. 3. Labels assigned by humans might be inconsistent or leave out the factors that are difficult for humans to identify".

Following these critiques and moving away from semi-supervision a number of weak-supervision frameworks emerged.

In "Weakly-Supervised Disentanglement Without Compromises," **?** model observations not as independent samples, but as pairs (x_1, x_2) that share at least one underlying factor of variation, even if the label of that factor is unknown.

For example, in a video of a moving arm, two adjacent frames might share the same "background color" and "object identity" but differ in "arm position." The weak label here is simply the knowledge that some factors are shared, without specifying what they are (unlike **?** who require the explicit label $y = $ "Digit 7").

They proved theoretically that knowing how many factors have changed (or stayed the same) between pairs is sufficient to guarantee disentanglement. This relaxes the requirement of **?** for expensive annotations for parts of the data while maintaining the capability of supervised disentanglement, only requiring a data collection process that provides paired samples (e.g., temporal coherence in video, or multi-camera setups).

? also critique the necessity for labels **?** and take a similar approach to **?**, but focus more on the geometric structure of the latent space, utilizing a ranking-based loss to ensure that pairwise similarities in the data are preserved as proximity in specific latent dimensions. While **?** provide formal identifiability guarantees using *rank-1* pairs, where exactly one factor is varied, **?** rely on more flexible binary same/different judgments. This allows the model to learn from supervision that indicates whether any factor is shared, without requiring the explicit knowledge of which specific factor differs between the pair.

? extend these principles to hand pose estimation and image synthesis, claiming the ability to learn interpretable disentangled representations without the necessity for additional weak labels. However, their framework still fundamentally relies on explicit pose annotations to supervise the partitioning of the latent space into pose and appearance components. This highlights a recurring theme in the literature where, despite claims of minimizing supervision, the underlying disentanglement mechanism remains dependent on the primary task's labels to maintain structural integrity bringing us back to the impossibility theorem from **?**.

4.4.2 Multimodal VAEs and Mutual Supervision

?, in "Learning Multimodal VAEs through Mutual Supervision," extended the semi-supervised framework to scenarios with multiple high-dimensional modalities (e.g., Image + Text).

In **?**, supervision comes from a low-dimensional label y. In Multimodal VAEs, supervision comes from another rich modality. **?** utilized a Product-of-Experts (PoE) aggregation. The joint posterior is approximated as the product of individual posteriors:

$$q(z|x_{\text{img}}, x_{\text{text}}) \propto q(z|x_{\text{img}}) \cdot q(z|x_{\text{text}})$$

This structure enforces a form of inter-modality disentanglement. Similar to **?**, the shared information (semantics) must be encoded in the intersection of the posteriors, while modality-specific noise is filtered out. This mutual supervision allows the model to learn robust representations even when one modality is missing, extending the partial-specification idea of **?** to complex, unstructured labels.

4.5 Applications and Integrations in the Literature

While multiple papers simply use SSVAE as a baseline in their experiments (**??**), other apply various aspects of **?**s framework to their method. **?** critiques the prevalent use of synthetic data in the literature, several papers apply SSVAE to real world problems. This section discusses several works that apply the concepts of SSVAEs or use it as a baseline in their benchmark, which provides perspective into the utility of the framework proposed by **?** in real world conditions.

4.5.1 Domain Adaptation in the Medical Domain

? apply the framework of SSVAEs to the critical problem of unsupervised domain adaptation in medical imaging (e.g., segmenting livers in CT vs. MRI scans). Building on the latent decomposition proposed by **?**, **?**s framework utilizes separate encoders (E_c content code and E_s style code) to partition representations into a shared content manifold for anatomical preservation and a domain-specific style subspace for modality features. This architecture facilitates many-to-many cross-domain translation via style-code swapping and adversarial alignment, forcing the content space to converge on a modality-agnostic representation suitable for unsupervised liver segmentation.

By enforcing that the content code is shared while the style code is domain-specific, they could train a segmenter on labeled CT scans and apply it to unlabeled MRI scans by swapping the style codes. This effectively treats the "domain" (CT/MRI) as the specified factor y in **?**'s framework, but extends the mechanism to allow for image-to-image translation via the disentangled latent space.

? extend the semi-supervised generative framework to the medical domain through the development of EVA, a model designed for the generation of longitudinal electronic health records. Their approach builds upon the SSVAE architecture by introducing hierarchically factorized latent variables to better capture the complex dependencies inherent in temporal patient data. In this hierarchical structure, the upper-level latent variables are shared across the entire population to represent common clinical characteristics and global trends, while the lower-level variables remain patient-specific to capture individual disease trajectories and variations. By incorporating this multi-level factorization, the model ensures semantic consistency across the population while simultaneously allowing for the fine-grained representation of individual patient histories.

? work on "multi-objective de novo drug design with conditional graph generative models" and mention SSVAEs as a possible extension to improve their work

4.5.2 Causal Extension and Robustness

? provide a causal extension to the graphical latent variable framework, applying it to the study of neural network robustness. By adopting a setup similar to the semi-supervised generative models in **?**, they frame the relationship between latent factors and observed data through a structural causal model (SCM). This causal lens allows for a formal investigation into how perturbations and distribution shifts propagate through the model's representations. Their work demonstrates that leveraging a graphical latent model is not only useful for disentanglement but is also critical for analyzing and improving the robustness of neural networks against adversarial or environmental interventions.

4.5.3 Catastrophic Forgetting and Generative Replay

To address catastrophic forgetting in deep learning models trained sequentially, **?** introduced the Lifelong VAEGAN, which combines the inference capabilities of VAEs with the high-quality generation of Generative Adversarial Networks (GANs). Instead of a standard teacher-student setup, the model employs a generative replay mechanism and a two-step "wake-dreaming" optimization. In the "wake" phase, the generator is updated to approximate the distribution of both current data and replayed samples from previous tasks. In the "dreaming" phase, the model maximizes the log-likelihood of these samples to learn shared, disentangled latent representations across multiple domains.
Building on **?**, **?** extend their Lifelong VAEGAN framework to the supervised setting and directly compare to SSVAE in their experiments where it has the second lowest classification error after the stacked generative semi-supervised model (M1+M2) model from **?**.
In a later paper, **?** extend the SSVAE framework into the domain of continual learning by integrating it within a lifelong teacher-student network. Their method builds upon the generative principles established by **?** to facilitate knowledge retention and transfer across sequential learning tasks, utilizing the original model as a primary baseline for experimental comparison. Notably, while the authors formally cite **?**, they attribute the architectural foundations to the **?** model in their technical descriptions, reflecting the shared heritage of these semi-supervised generative approaches. This extension demonstrates how the structured latent spaces of SSVAEs can be leveraged to mitigate catastrophic forgetting in dynamic data environments. Further, they find that the SSVAE architecture performs well in an experiment comparing the semi-supervised classification results on MNIST data, when considering MNIST to MNIST-Fashion lifelong learning. Specifically, they compare ten architectures of which SSVAE as the second lowest error.

5   Experiments

5.1   Implementation Details

Following previous work, our experiments where conducted with the MNIST dataset **?** using the digit label for partial supervision. We use the same fully connected feed forward architecture as **?**, with Rectified Linear Units (ReLU) **?** activations. The SSVAE architecture used in the following experiments is detailed in Figure 1.
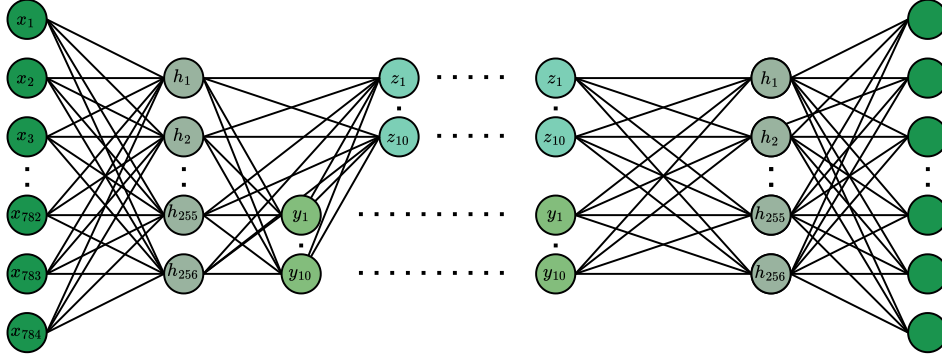

Figure 1: SSVAE architecture used throughout all experiments. The network has $28*28 = 784$ inputs matching the MNIST input, 256 hidden neurons, 10 style variables and 10 digit neurons.

Following **?** we use the Adam optimizer (**?**) with default parameters, a learning rate of 10^{-3} and a batch size of 128. Unlike the original paper we train for 40 instead of 200 epochs due computational constraints. As shown in Appendix B this is sufficient for all trained models to converge. Each experiment is repeated with 10 different random seeds. Experiments were run on a Nvidia Tesla P40 GPU and implemented with the ProbTorch package introduced by **?**. To reproduce the experiments presented in this paper see the code at: `https://github.com/davidbhoffmann/ssvae`.

5.2   Supervision Weight

While **?** explore the supervision rate ρ and therefore indirectly γ, the supervision weight α is not explored by the authors. Furthermore, the authors make conflicting statements about the value of α in the MNIST experiments as critiqued in subsection 4.1. In the Yale-B experiment alpha is set to a different value without justification and in the Multi-MNIST experiments the value of α is never mentioned. While **?** change its value, indicating its importance, they leave the reader guessing at the effect of α.

To address this, we explore how robust the SSVAE framework is with respect to the supervision weight α which we evaluate for eight different values: 0.1, 0.5, 1, 10, 25, 50, 75 and 100. The remaining configurations follow the setup explained in subsection 5.1.
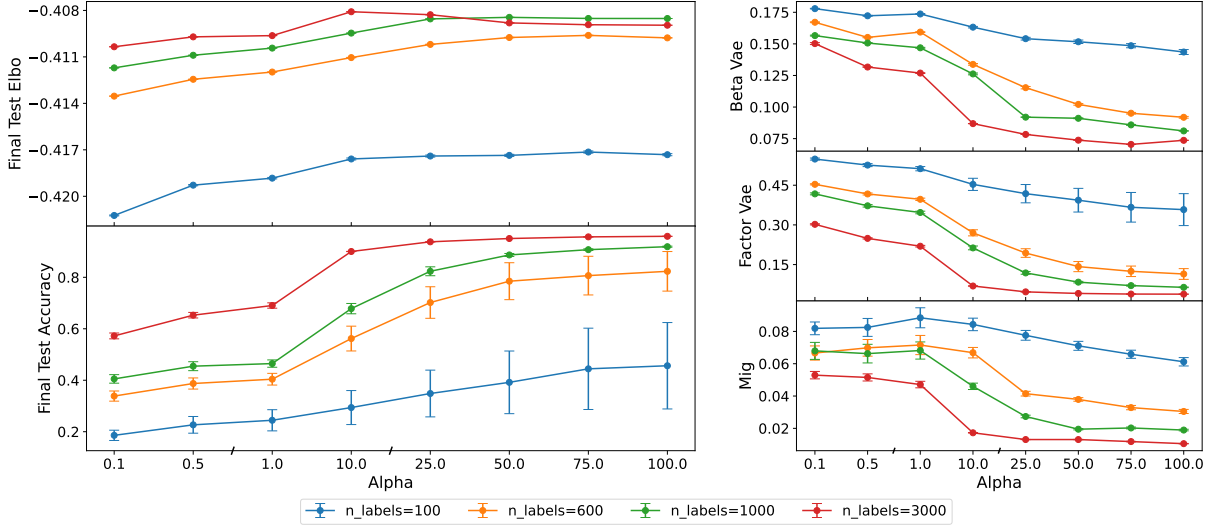
Figure 2: **Supervision Weight** experiment results with varying α for supervised set sizes of 100, 600, 1000 and 3000. Error bars indicate variance across 10 random seeds. **Top Left:** final test ELBO for varying α remains relatively stable. **Bottom Left:** final test accuracy of y improves for higher values of α. **Right:** disentanglement metrics (β-VAE **?**, Factor-VAE **?**, Mutual Information Gap **?**) decreases with increasing α across all three metrics.

Figure 2 shows that classification accuracy of the semi-supervised y variable is sensitive to the supervision weight α. As expected increasing α leads to better classification performance, although the effect of increasing the value beyond 10 only leads to marginal improvements. While we did not report final training accuracy in the experiments, we can still observe there is no overfitting for high α values to the point of decreasing test accuracy.

The final test ELBO remains relatively constant with respect to α, which can also be seen in Appendix B. A minor increase of the ELBO can be observed wit increasing α. The results further show that for the unsupervised z, stronger supervision decreases disentanglement.

5.3 Label Corruption

The most frequent critique identified in our review process was the use of labels (**????**) as explained in subsubsection 4.4.1. One aspect of these critiques is the potential for incorrect human annotations.

To investigate the sensitivity of SSVAEs to potentially noisy labels, we use the setup as explained above and introduce a random corrupt labels into the training process. Specifically, in each epoch a specified amount of labels is flipped to a random class. We evaluate the SSVAEs for corruption rates of 0.0, 0.01, 0.1, 0.2, 0.5 and 1. The remaining configurations follow the setup explained in subsection 5.1.

The result of this experiment is shown in Figure 3. As expected We find that classification accuracy of the latent y variable decreases as the corruption rate of the labels increases.
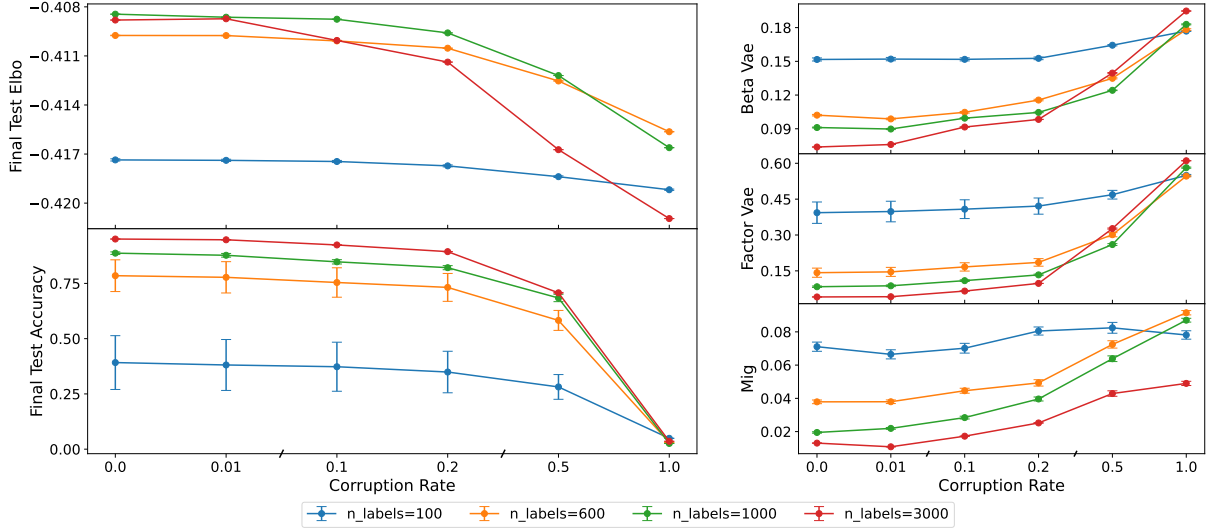
Figure 3: **Label Corruption** experiment results with varying label corruption rate for supervised set sizes of 100, 600, 1000 and 3000. Error bars indicate variance across 10 random seeds. **Top Left:** final test ELBO for varying noise levels remains relatively stable. **Bottom Left:** while the final test accuracy of y decreases for higher corruption rates it remains relatively high for up to 20% noise. **Right:** disentanglement (as measured through β-VAE **?**, Factor-VAE **?**, Mutual Information Gap **?**) increase with a higher corruption rate across all three metrics.

However, the accuracy does remain relatively stable up to a corruption rate of 20%. Similar to the the supervision weight experiments, the final test ELBO remains relatively constant. However, a marginal ELBO drop with increased corruption rates is noticeable across configurations. The results further show that for the unsupervised z, more label corruption in y decreases disentanglement.

6 Discussion

In this work, we presented a unified framework for learning disentangled representations by integrating deep generative models with partially-specified probabilistic graphical models. Unlike standard variational autoencoders that often assume a flat latent space, our approach allows for the injection of domain knowledge through structured dependencies while retaining the flexibility of neural networks for function approximation. By mapping these definitions onto stochastic computation graphs, we demonstrated that end-to-end semi-supervised learning is feasible even with complex, non-conjugate dependency structures.

6.1 Performance and Compositionality

$[cite_s tart] The experimental results across MNIST, SVHN, and Multi-MNIST confirm that imposing stru$ $264, 271].[cite_s tart] In the classification tasks, our method performed on par with established baselines, der$ $sampled objective is an effective estimator for semi-supervised learning in structured models [cite:$ $271, 152]. Most notably, the Multi-MNIST experiments highlighted an important trade-$ $off between raw performance and model modularity. [cite_s tart] While the ``flat'' model achieved superior co$ $trained MNIST model | demonstrated the framework's capacity for compositional design [cite:$ $376, 381].[cite_s tart] Although the nested model incurred a performance cost in counting accuracy (84.79\%), i$ $to-end retraining of the component models [cite: 381, 383].$

6.2 The Role of Structure in Disentanglement

A key finding of this study is that partial supervision, when combined with an explicit graphical structure, effectively enforces disentanglement where purely unsupervised methods often fail. $[cite_s tart] By defining specific nodes for interpretable factors | such as the digit class or the n$ $169, 170].[cite_s tart] This was evident in the SVHN and Intrinsic Faces experiments, where the model success$ $282, 286].[cite_s tart] This validates the hypothesis that ``partially-specified'' models occupy a sweet spot betw$ $specified graphical models (which require expensive engineering) and unstructured deep generative model$ $30, 38, 45].$

6.3 Limitations in Heterogeneous Sparse Supervision

While the framework generalizes to arbitrary dependency structures, its practical application in real-world settings with multiple partially observed variables reveals significant limitations. $[cite_s tart] The current formulation relies on a binary distinction between unsupervised data D and s$ assuming a consistent set of available labels for the supervised portion[cite: 116]. However, in complex real-world datasets involving K distinct supervised variables $\mathbf{y} = \{y_1, \ldots, y_K\}$, supervision is often heterogeneous; a datapoint may possess annotations for a random subset of variables while others are missing. $[cite_s tart] Under the proposed framework, variables are treated as$ $179].Implementing this dynamically in a batched setting becomes non-trivial, as the objective function f$ potential observation patterns. This combinatorial explosion necessitates a bespoke and brittle implementation of the stochastic computation graph, undermining the flexibility intended by the general framework.

6.4   Future Directions

[cite_start]The current implementation relies on static graph definitions, but the natural evolution of this wo
390]. [cite_start]Probabilistic programs would allow for more expressive models involving recursion and con
391]. [cite_start]Future work will explore amortizing inference over these dynamic structures, enabling the l
392, 393].

7   Conclusion

A concise summary of contents and results

The trajectory of research stemming from "Learning Disentangled Representations with Semi-Supervised Deep Generative Models" illustrates the rigorous self-correction mechanism of the scientific community. Siddharth et al. (2017) correctly identified that supervision was the key to unlocking interpretable representations, effectively predicting the impossibility results that would later prove purely unsupervised disentanglement futile.

However, the specific mechanism they proposed—a simple partitioning of the latent space into label y and style z—proved to be structurally insufficient for complex real-world data. It suffered from semantic conflation (Joy et al.), lacked theoretical guarantees for the unspecified variables (Locatello et al.), and was prone to optimization failures (Feng et al., Mattei et al.).

The extensions detailed in this report represent a maturation of the field. We have moved from simple semi-supervision to weakly supervised causal mechanisms, from static models to lifelong learning agents, and from deterministic encoders to uncertainty-aware Bayesian frameworks. The legacy of Siddharth et al. lies not just in their specific architecture, but in shifting the paradigm towards principled, supervision-guided structure learning.

A Systematic Review Process

To identify relevant extension and limitations of **?** in the literature, the results of several retrieval methods where synthesized. Note that the papers identified through the different methods had some overlap. For any paper to be considered it had fulfil one of the following selection criteria: (1) directly critique **?**, (2) cite the paper in the context of a more general critique, (3) extend the SSVAE framework or (4) use it in an experiment.

Firstly, Connected Papers (`https://www.connectedpapers.com`) lead to a list of 10 derivative works, of which only one matched the selection criteria.

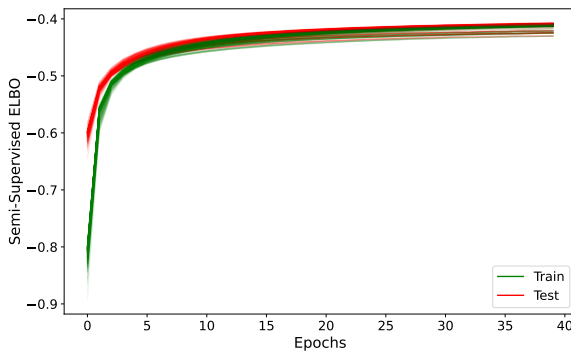Secondly, Google Scholar (`https://scholar.google.com`) was used to find results relevant literature in the list of 445 papers which cite **?**. The first 150 citing papers sorted by relevance where checked for the selection criteria. For the first 80 results, all paragraphs that mentioned **?** where reviewed in detail and for the remaining 70 papers only papers that mentioned one of the following keywords in the title where considered: *semi-supervised*, *weak-supervision* and *label*. With this method 21 relevant papers where identified.

Additionally, the search functionality of Google Scholar was used to find results for the search terms *semi-supervised disentanglement* and *supervised disentanglement*. For each term the first 10 results where reviewed in detail leading to a total of 4 relevant papers.
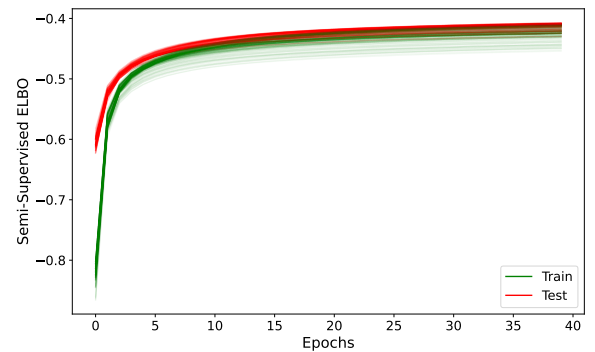
Lastly, Gemini-3-Pro was used to generate a report with relevant extensions and critiques which lead to 5 relevant papers.

B SSVAE Training

Figure 4a and Figure 4b show the training curve plots of all runs from both the supervision factor (alpha) and corruption rate experiments. The figures show that all runs converge without overfitting.

(a) All training runs of the **supervision factor α experiment** in subsection 5.2

(b) All training runs of the **corruption rate experiment** in subsection 5.3

Declaration of authorship

I hereby declare that the report submitted is my own unaided work. All direct or indirect sources used are acknowledged as references. I am aware that the Thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the report as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future Theses submitted. Further rights of reproduction and usage, however, are not granted here. This paper was not previously presented to another examination board and has not been published.

Location, date

Name