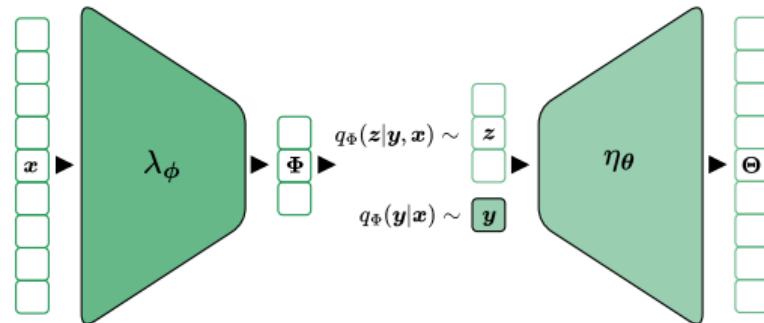


Learning Disentangled Representations with Semi-Supervised Deep Generative Models

David B. Hoffmann

January 23, 2026



Outline

- 1 Recap: Semi-Supervised Disentanglement
- 2 Own Critiques
- 3 Literature Review: Limitations and Extensions
- 4 Experiments
- 5 Conclusion

Fully specified and supervised graphical models which are interpretable.

Unsupervised variational auto-encoders which are uninterpretable.

Fully specified and supervised graphical models which are interpretable.



GAP

A diagram consisting of two light green rectangular boxes. The left box contains the text "Fully specified and supervised graphical models which are interpretable.". The right box contains the text "Unsupervised variational auto-encoders which are uninterpretable.". Two black arrows point from each text box towards the center, where the word "GAP" is written in large, bold, black capital letters.

Unsupervised variational auto-encoders which are uninterpretable.

Fully specified and supervised graphical models which are interpretable.



Semi-Supervised VAE [24]

Unsupervised variational auto-encoders which are uninterpretable.

Formulation

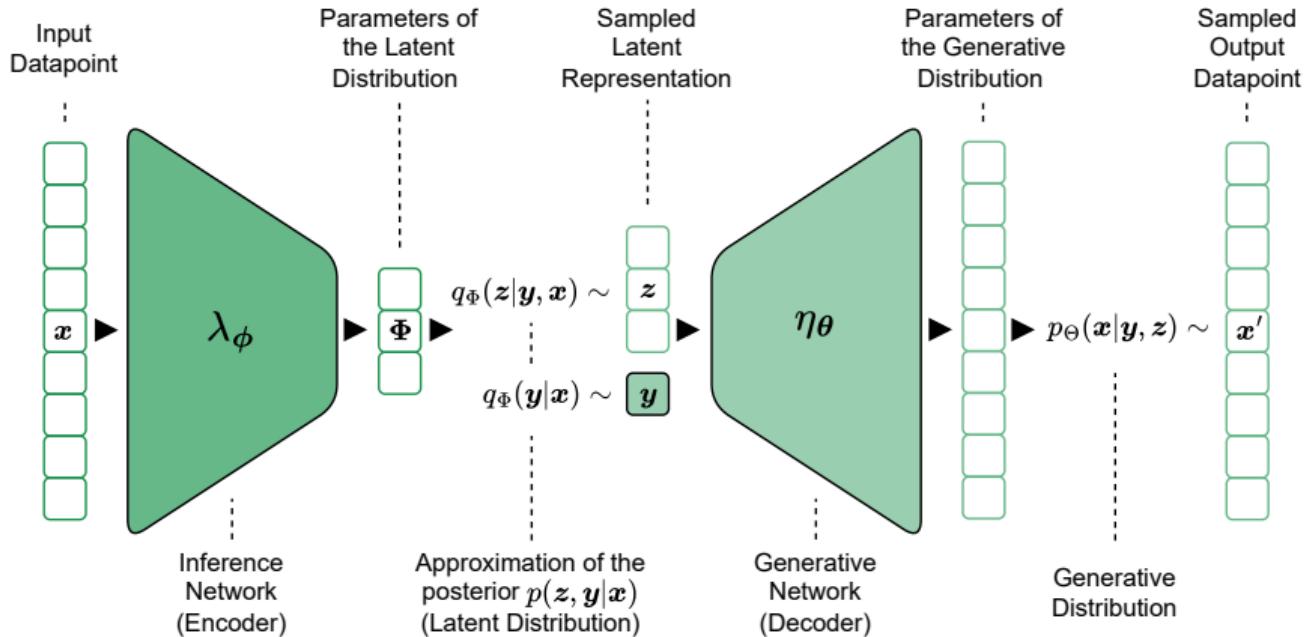


Figure: Formulation of the semi-supervised disentanglement framework

Formulation

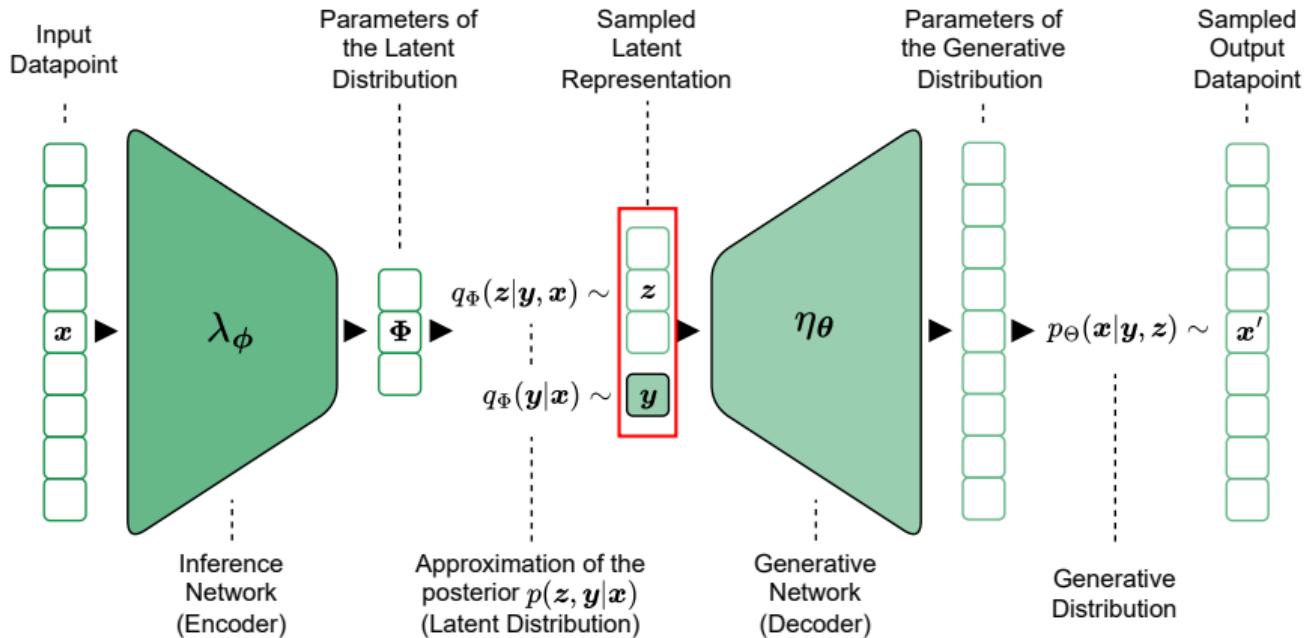


Figure: Formulation of the semi-supervised disentanglement framework

Semi-Supervised Objective Formulation

The total objective combines unsupervised (x) and supervised (x, y) data, weighted by γ :

$$\mathcal{L}(\theta, \phi) = \sum_{x \in \mathcal{D}} \mathcal{L}(\theta, \phi; x) + \gamma \sum_{(x, y) \in \mathcal{D}^{\text{sup}}} \mathcal{L}_{\text{sup}}(\theta, \phi; x, y)$$

Semi-Supervised Objective Formulation

The total objective combines unsupervised (x) and supervised ((x, y)) data, weighted by γ :

$$\mathcal{L}(\theta, \phi) = \sum_{x \in \mathcal{D}} \mathcal{L}(\theta, \phi; x) + \gamma \sum_{(x, y) \in \mathcal{D}^{\text{sup}}} \mathcal{L}_{\text{sup}}(\theta, \phi; x, y)$$

The Supervised Term \mathcal{L}_{sup} : Defined to jointly maximize the generative likelihood and discriminative power:

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

Semi-Supervised Objective Formulation

The total objective combines unsupervised (x) and supervised ((x, y)) data, weighted by γ :

$$\mathcal{L}(\theta, \phi) = \sum_{x \in \mathcal{D}} \mathcal{L}(\theta, \phi; x) + \gamma \sum_{(x, y) \in \mathcal{D}^{\text{sup}}} \mathcal{L}_{\text{sup}}(\theta, \phi; x, y)$$

The Supervised Term \mathcal{L}_{sup} : Defined to jointly maximize the generative likelihood and discriminative power:

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

- To allow for arbitrary dependencies in the encoder, importance sampling with proposals from the unconditioned encoder $q_\phi(z|x)$, is used.
- The discriminative term is also intractable, but bounded by the same weights:
 $\log q_\phi(y|x) \geq \log \left(\frac{1}{S} \sum_s w_s \right) \approx \log \hat{Z}$

Outline

- 1 Recap: Semi-Supervised Disentanglement
- 2 Own Critiques
- 3 Literature Review: Limitations and Extensions
- 4 Experiments
- 5 Conclusion

Own Critiques: Presentation Issues

While the framework is technically correct, several presentation issues disrupt reading flow:

While the framework is technically correct, several presentation issues disrupt reading flow:

- **Incomplete Figure Documentation**

- ▶ Some figures lack captions and are never referenced in text.
- ▶ Other figures are referenced without interpretation guidance.
- ▶ Example: Shading in computational graph nodes is never explained.

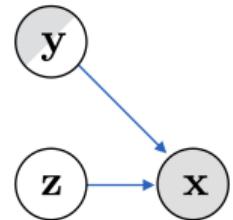


Figure: MNIST graph from the paper.

While the framework is technically correct, several presentation issues disrupt reading flow:

- **Incomplete Figure Documentation**

- ▶ Some figures lack captions and are never referenced in text.
- ▶ Other figures are referenced without interpretation guidance.
- ▶ Example: Shading in computational graph nodes is never explained.

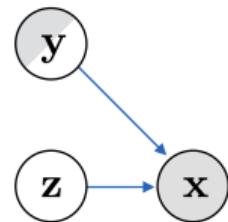


Figure: MNIST graph from the paper.

- **Undefined Variables**

- ▶ Example: Variable n in the YaleB dataset [9] experiment relationship is never defined.

● Factorisable Generalization

- ▶ Authors claim that "*a representation that has some factorisable structure, and consistent semantics associated to different parts, is more likely to generalise to a new task*", as one of the selling points of their method.
- ▶ They do not provide any theoretical or empirical evidence for this claim.
- ▶ Locatello et al. (2018) [21] investigate this, but find no evidence to support the claim.

● Factorisable Generalization

- ▶ Authors claim that "*a representation that has some factorisable structure, and consistent semantics associated to different parts, is more likely to generalise to a new task*", as one of the selling points of their method.
- ▶ They do not provide any theoretical or empirical evidence for this claim.
- ▶ Locatello et al. (2018) [21] investigate this, but find no evidence to support the claim.

● Learning Relationships

- ▶ In the YaleB experiment, the authors claim to "demonstrate that [their] generative model still learns the correct relationship over [the] latent variables" referring to this relationship:
$$(n \cdot l) \times r + \epsilon$$
- ▶ While they qualitatively demonstrate generative capacity as well as classification and regression performance, it is not shown that the model has learned this specific relationship.

Outline

- 1 Recap: Semi-Supervised Disentanglement
- 2 Own Critiques
- 3 Literature Review: Limitations and Extensions
- 4 Experiments
- 5 Conclusion

- **Inclusion Criteria:**

- ▶ Directly critique Narayanaswamy et al. (2017),
- ▶ cite the paper in the context of a more general critique,
- ▶ extend the Semi-Supervised Variational Autoencoder (SSVAE) framework, or
- ▶ use it in an experiment.

- **Inclusion Criteria:**

- ▶ Directly critique Narayanaswamy et al. (2017),
- ▶ cite the paper in the context of a more general critique,
- ▶ extend the SSVAE framework, or
- ▶ use it in an experiment.

- **Sources:**

- ▶ Review all derivative works found on Connected Papers (1 match).
- ▶ Review first 150 Google Scholar citations sorted by relevance (21 matches).
- ▶ Review first 10 results of Google Scholar Search for "semi-supervised disentanglement" and "supervised disentanglement" (4 matches).
- ▶ Gemini-3-Pro research report for "semi-supervised disentanglement" (5 matches).

- **Inclusion Criteria:**

- ▶ Directly critique Narayanaswamy et al. (2017),
- ▶ cite the paper in the context of a more general critique,
- ▶ extend the SSVAE framework, or
- ▶ use it in an experiment.

- **Sources:**

- ▶ Review all derivative works found on Connected Papers (1 match).
- ▶ Review first 150 Google Scholar citations sorted by relevance (21 matches).
- ▶ Review first 10 results of Google Scholar Search for "semi-supervised disentanglement" and "supervised disentanglement" (4 matches).
- ▶ Gemini-3-Pro research report for "semi-supervised disentanglement" (5 matches).

- **Final Set:** removing duplicates results in a final set of 26 papers. Note: Most work related to SSVAEs cites [17] (with 4080 citations) while the generalization by [24] is only cited 446.

- **General Critiques:** Including the impossibility theorem by Locatello et al. [21], critiques of isotropic prior assumption [4, 8] and unbounded likelihoods [23].

- **General Critiques:** Including the impossibility theorem by Locatello et al. [21], critiques of isotropic prior assumption [4, 8] and unbounded likelihoods [23].
- **Applications of SSVAEs:** While some papers simply use SSVAEs as a baseline [10, 18], others apply them to specific domains such as healthcare [28, 2, 19], causality [32] or continual learning [30, 31].

- **General Critiques:** Including the impossibility theorem by Locatello et al. [21], critiques of isotropic prior assumption [4, 8] and unbounded likelihoods [23].
- **Applications of SSVAEs:** While some papers simply use SSVAEs as a baseline [10, 18], others apply them to specific domains such as healthcare [28, 2, 19], causality [32] or continual learning [30, 31].
- **Effects of Supervising Disentanglement:** Several papers question whether the SSVAE latent space is truly disentangled [12, 25, 27].

- **General Critiques:** Including the impossibility theorem by Locatello et al. [21], critiques of isotropic prior assumption [4, 8] and unbounded likelihoods [23].
- **Applications of SSVAEs:** While some papers simply use SSVAEs as a baseline [10, 18], others apply them to specific domains such as healthcare [28, 2, 19], causality [32] or continual learning [30, 31].
- **Effects of Supervising Disentanglement:** Several papers question whether the SSVAE latent space is truly disentangled [12, 25, 27].
- **From Semi- to Weak Supervision:** Critique of costly labels [20, 3, 14, 15] and the impossibility theorem [21] lead to weakly-supervised approaches [22, 5, 29, 13].

Critique of Narayanaswamy et al. [24]'s Label-Conditioned Decoder from Joy et al. [12].

- **Assumption:** Decoder: $p(x | y, z)$ learns rigid separation where label y handles class info and latent z handles rest.

Critique of Narayanaswamy et al. [24]'s Label-Conditioned Decoder from Joy et al. [12].

- **Assumption:** Decoder: $p(x | y, z)$ learns rigid separation where label y handles class info and latent z handles rest.
- **Flaw:** Labels often imply continuous characteristics which break this assumption.

Critique of Narayanaswamy et al. [24]'s Label-Conditioned Decoder from Joy et al. [12].

- **Assumption:** Decoder: $p(x | y, z)$ learns rigid separation where label y handles class info and latent z handles rest.
- **Flaw:** Labels often imply continuous characteristics which break this assumption.
- **Result:** Model learns the label value instead of the rich semantics associated with it.

Critique of Narayanaswamy et al. [24]'s Label-Conditioned Decoder from Joy et al. [12].

- **Assumption:** Decoder: $p(x | y, z)$ learns rigid separation where label y handles class info and latent z handles rest.
- **Flaw:** Labels often imply continuous characteristics which break this assumption.
- **Result:** Model learns the label value instead of the rich semantics associated with it.
- Joy et al. term this **semantic conflation** and argue that labels are "actively harmful" to disentanglement.

Critique of Narayanaswamy et al. [24]'s Label-Conditioned Decoder from Joy et al. [12].

- **Assumption:** Decoder: $p(x | y, z)$ learns rigid separation where label y handles class info and latent z handles rest.
- **Flaw:** Labels often imply continuous characteristics which break this assumption.
- **Result:** Model learns the label value instead of the rich semantics associated with it.
- Joy et al. term this **semantic conflation** and argue that labels are "actively harmful" to disentanglement.

Solution

- Joy et al. [12] propose the CCVAE.
- Split latent space: z_c (characteristic) + z_s (salient).
- z_c is indirectly supervised through an auxiliary classifier, forcing it into the continuous space.

From Semi- to Weak Supervision: Problem

- **Impossibility Theorem [21]:** Pure unsupervised disentanglement is impossible without inductive bias.

- **Impossibility Theorem [21]:** Pure unsupervised disentanglement is impossible without inductive bias.
- **Critiques of Labels** by [20, 3, 14, 15].
 - ▶ Labels are costly (require human annotation).
 - ▶ Labels may be inconsistent or incomplete.
 - ▶ Humans learn factors of variation unsupervised [26].

- **Impossibility Theorem [21]:** Pure unsupervised disentanglement is impossible without inductive bias.
- **Critiques of Labels** by [20, 3, 14, 15].
 - ▶ Labels are costly (require human annotation).
 - ▶ Labels may be inconsistent or incomplete.
 - ▶ Humans learn factors of variation unsupervised [26].
- **Implication:** find a middle ground which does not require full supervision but still leads to disentanglement.

- **Impossibility Theorem [21]:** Pure unsupervised disentanglement is impossible without inductive bias.
- **Critiques of Labels** by [20, 3, 14, 15].
 - ▶ Labels are costly (require human annotation).
 - ▶ Labels may be inconsistent or incomplete.
 - ▶ Humans learn factors of variation unsupervised [26].
- **Implication:** find a middle ground which does not require full supervision but still leads to disentanglement.

Solution

Locatello et al. [22] suggest learning from Paired Observations.

- Key Idea: Use pairs (x_1, x_2) that share underlying factors.
- [22] Prove that knowing *how many* factors changed is sufficient for guarantee disentanglement.
- No need to know which factors changed.

Outline

- 1 Recap: Semi-Supervised Disentanglement
- 2 Own Critiques
- 3 Literature Review: Limitations and Extensions
- 4 Experiments
- 5 Conclusion

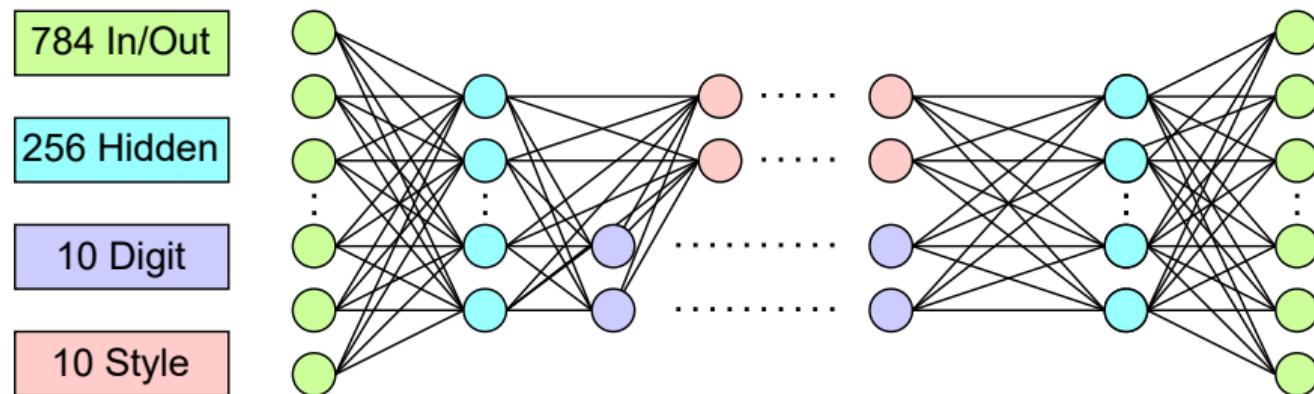
- **Dataset:** MNIST [7]

- ▶ Digit label used for partial supervision.
- ▶ Input: $28 \times 28 = 784$ pixels.

- **Dataset:** MNIST [7]

- ▶ Digit label used for partial supervision.
- ▶ Input: $28 \times 28 = 784$ pixels.

- **Architecture:** Linear network with ReLU activations [1] (same as [24]).



- **Training Configuration:**

- ▶ Optimizer: Adam [16] with default parameters.
- ▶ Learning rate: 10^{-3} , Batch size: 128.
- ▶ Epochs: 40 (vs. 200 in original paper due to computational constraints).
- ▶ 10 random seeds per configuration.
- ▶ Supervised set sizes of 100, 600, 1000 and 3000 labels per configuration.

- **Training Configuration:**

- ▶ Optimizer: Adam [16] with default parameters.
- ▶ Learning rate: 10^{-3} , Batch size: 128.
- ▶ Epochs: 40 (vs. 200 in original paper due to computational constraints).
- ▶ 10 random seeds per configuration.
- ▶ Supervised set sizes of 100, 600, 1000 and 3000 labels per configuration.

- **Implementation:**

- ▶ Model implementation with ProbTorch [24].
- ▶ 560 models trained on a Nvidia Tesla P40 GPU.
- ▶ Code available at: <https://github.com/davidbhoffmann/ssvae> (repository will be made public upon final submission).

Experiment: Label Corruption

- **Motivation:** The most frequent critique of SSVAE is the use of expensive potentially noisy labels [20, 3, 14, 15].

Experiment: Label Corruption

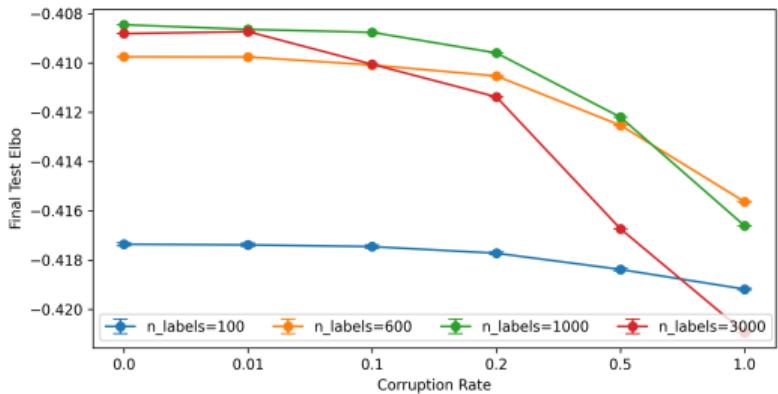
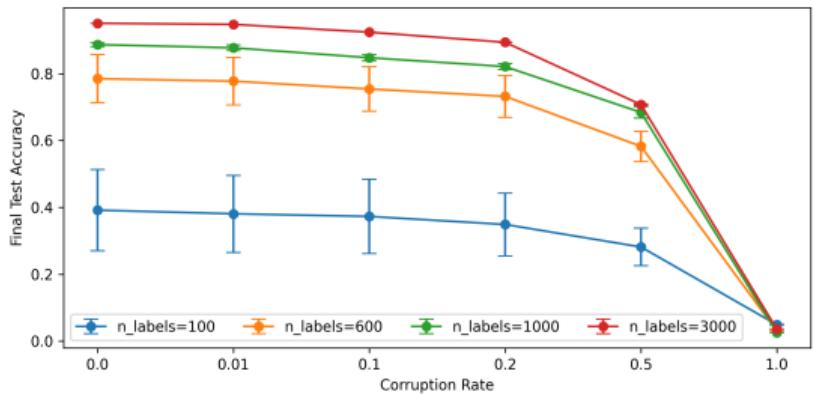
- **Motivation:** The most frequent critique of SSVAE is the use of expensive potentially noisy labels [20, 3, 14, 15].
- **Research Question:** How robust is the variable specification in SSVAE to label noise?

- **Motivation:** The most frequent critique of SSVAE is the use of expensive potentially noisy labels [20, 3, 14, 15].
- **Research Question:** How robust is the variable specification in SSVAE to label noise?
- **Setup:**
 - ▶ Randomly corrupt 0% to 100% of labels (0%, 1%, 10%, 20%, 50%, 100%).
 - ▶ Supervision weight is fixed to $\alpha = 50$ (as in [24]).
 - ▶ Measure disentanglement of z with: Beta-VAE metric, Factor-VAE metric, Mutual Information Gap (MIG) [6].
 - ▶ Measure accuracy of the supervised variable y (MNIST digit label).

Experiment: Label Corruption

Accuracy and ELBO Results:

- Test accuracy stable up to $\approx 20\%$ corruption for MNIST digit label.
- As corruption increases, accuracy drops below random chance (10%).
- ELBO drops slightly with increasing corruption, but overall remains stable.



Experiment: Label Corruption

Effect on Disentanglement Scores:

- Disentanglement is higher for smaller supervised set sizes.
- Further, scores increase with label corruption. More pronounced for larger supervised set sizes.
- Indicates that supervision harms disentanglement of the style latent z .

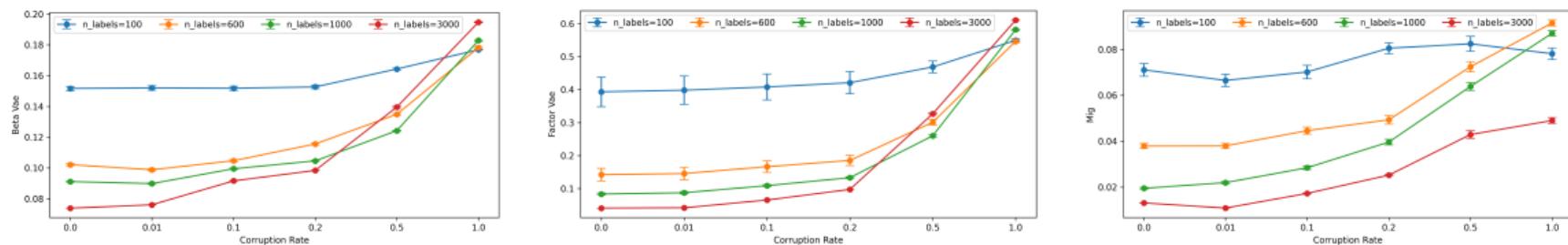


Figure: Disentanglement score (left to right: Beta-VAE, Factor-VAE and MIG) for varying label corruption rates.

Experiment: Supervision Weight Sensitivity

- **Motivation:** While Narayanaswamy et al. [24] investigate the effect of γ , the supervision weight α is not explored. Given the semantic conflation argument from Joy et al. [12] we want to further explore the effects of supervision.

$$\mathcal{L}(\theta, \phi) = \sum_{x \in \mathcal{D}} \mathcal{L}(\theta, \phi; x) + \gamma \sum_{(x,y) \in \mathcal{D}^{\text{sup}}} \mathbb{E}_{q_\phi(z|x,y)} \left[\log \frac{p_\theta(x,y,z)}{q_\phi(z|x,y)} \right] + \alpha \log q_\phi(y|x)$$

Experiment: Supervision Weight Sensitivity

- **Motivation:** While Narayanaswamy et al. [24] investigate the effect of γ , the supervision weight α is not explored. Given the semantic conflation argument from Joy et al. [12] we want to further explore the effects of supervision.

$$\mathcal{L}(\theta, \phi) = \sum_{x \in \mathcal{D}} \mathcal{L}(\theta, \phi; x) + \gamma \sum_{(x,y) \in \mathcal{D}^{\text{sup}}} \mathbb{E}_{q_\phi(z|x,y)} \left[\log \frac{p_\theta(x,y,z)}{q_\phi(z|x,y)} \right] + \alpha \log q_\phi(y|x)$$

- **Research Question:** Does stronger supervision harm disentanglement?

Experiment: Supervision Weight Sensitivity

- **Motivation:** While Narayanaswamy et al. [24] investigate the effect of γ , the supervision weight α is not explored. Given the semantic conflation argument from Joy et al. [12] we want to further explore the effects of supervision.

$$\mathcal{L}(\theta, \phi) = \sum_{x \in \mathcal{D}} \mathcal{L}(\theta, \phi; x) + \gamma \sum_{(x,y) \in \mathcal{D}^{\text{sup}}} \mathbb{E}_{q_\phi(z|x,y)} \left[\log \frac{p_\theta(x,y,z)}{q_\phi(z|x,y)} \right] + \alpha \log q_\phi(y|x)$$

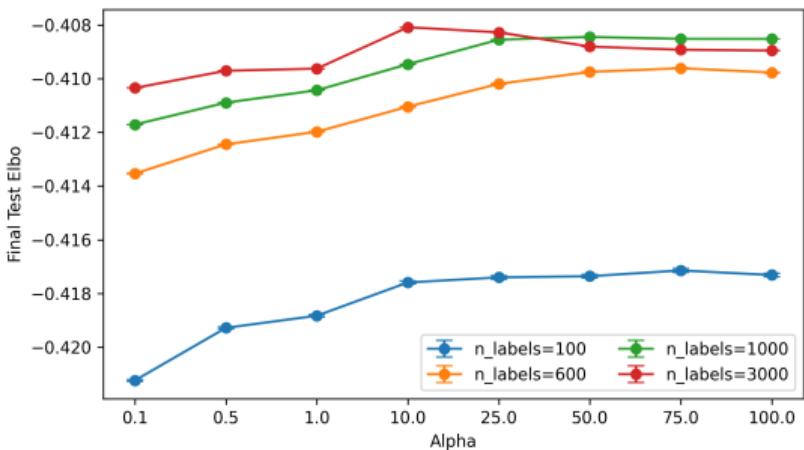
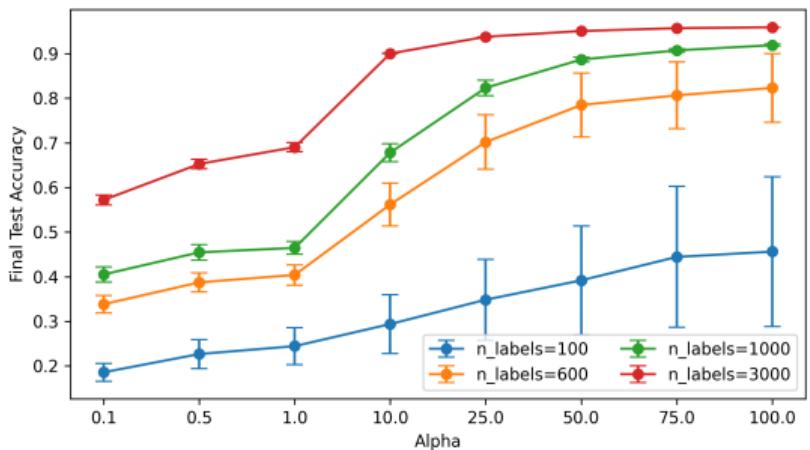
- **Research Question:** Does stronger supervision harm disentanglement?
- **Setup:**

- ▶ Supervision weight α varies from 0.1 to 100 (0.1, 0.5, 1, 10, 25, 50, 75, 100).
- ▶ Measure disentanglement of z with: Beta-VAE metric [11], Factor-VAE metric [15], Mutual Information Gap (MIG) [6].
- ▶ Measure accuracy of the supervised variable y (MNIST digit label).

Experiment: Supervision Weight Sensitivity

Accuracy and ELBO Results:

- Accuracy increases with the supervision weight α , while ELBO increases slightly but stays overall stable. Accuracy variance increases for smaller supervised set sizes and larger α .
- We don't observe overfitting even for large α values, possibly due to a low number of training epochs.



Experiment: Supervision Weight Sensitivity

Effect on Disentanglement Scores:

- Disentanglement is higher for smaller supervised set sizes.
- Further, scores decrease with larger supervision weights. More pronounced for larger supervised set sizes.
- Indicates a trade-off between supervision of y and disentanglement of z .

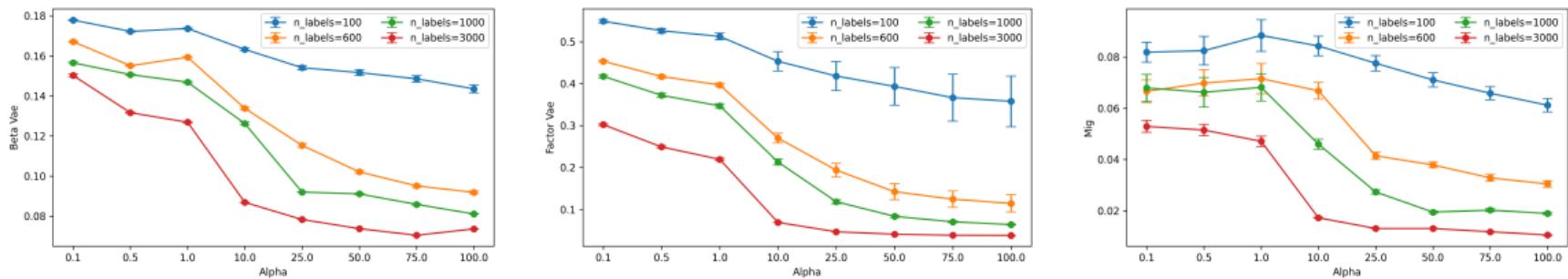


Figure: Disentanglement score (left to right: Beta-VAE, Factor-VAE and MIG) for varying supervision weights α .

Outline

- 1 Recap: Semi-Supervised Disentanglement
- 2 Own Critiques
- 3 Literature Review: Limitations and Extensions
- 4 Experiments
- 5 Conclusion

Conclusion

The SSVAE framework by Narayanaswamy et al. [24] provides a method for semi-supervised disentanglement of the latent space.

The SSVAE framework by Narayanaswamy et al. [24] provides a method for semi-supervised disentanglement of the latent space.

Literature Review

- Impossibility theorem motivates supervision [21].
- Applications of SSVAEs to healthcare [28, 2, 19], causality [32] or continual learning [30, 31].
- Critiques of disentanglement in SSVAE [12, 25, 27].
- Critiques of costly labels [20, 3, 14, 15] lead to weakly-supervised approaches [22, 5, 29, 13].

The SSVAE framework by Narayanaswamy et al. [24] provides a method for semi-supervised disentanglement of the latent space.

Literature Review

- Impossibility theorem motivates supervision [21].
- Applications of SSVAEs to healthcare [28, 2, 19], causality [32] or continual learning [30, 31].
- Critiques of disentanglement in SSVAE [12, 25, 27].
- Critiques of costly labels [20, 3, 14, 15] lead to weakly-supervised approaches [22, 5, 29, 13].

Experimental Results

- Label corruption experiment shows that SSVAE reverts to unsupervised VAE for high noise levels.
- Supervision weight sensitivity experiment reveals negative correlation between supervision strength and disentanglement quality.
- Both experiments validate critiques from literature review.

References I

- [1] Abien Fred Agarap. Deep learning using rectified linear units (reli). *ArXiv Preprint*, 2018.
- [2] Siddharth Biswal, Soumya Ghosh, Jon Duke, Bradley Malin, Walter Stewart, Cao Xiao, and Jimeng Sun. Eva: Generating longitudinal electronic health records using conditional variational autoencoders. In *Machine Learning for Healthcare Conference*, pages 260–282. PMLR, 2021.
- [3] Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [4] Francesco Paolo Casale, Adrian V. Dalca, Luca Saglietti, Jennifer Listgarten, and Nicoló Fusi. Gaussian process prior variational autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS 2018)*, 2018.
- [5] Junxiang Chen and Kayhan Batmanghelich. Weakly supervised disentanglement by pairwise similarities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3495–3502, 2020.

- [6] Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018.
- [7] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.
- [8] Babak Esmaeili, Hao Wu, Sarthak Jain, Alican Bozkurt, Narayanaswamy Siddharth, Brooks Paige, Dana H Brooks, Jennifer Dy, and Jan-Willem Meent. Structured disentangled representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2525–2534. PMLR, 2019.
- [9] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 23(6):643–660, 2001.

References III

- [10] Jonathan Gordon, John Bronskill, Matthias Bauer, Sebastian Nowozin, and Richard E. Turner. Meta-learning probabilistic inference for prediction. In *International Conference on Learning Representations (ICLR)*, 2019.
- [11] Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.
- [12] Tom Joy, Sebastian M. Schmon, Philip H. S. Torr, N. Siddharth, and Tom Rainforth. Capturing label characteristics in vaes. In *International Conference on Learning Representations (ICLR)*, 2021.
- [13] Tom Joy, Yuge Shi, Philip H.S. Torr, Tom Rainforth, Sebastian M. Schmon, and Narayanaswamy Siddharth. Learning multimodal vaes through mutual supervision. *ICLR*, 2022.

References IV

- [14] Qiao Ke, Xinhui Jing, Marcin Woźniak, Shuang Xu, Yunji Liang, and Jiangbin Zheng. Apgvae: Adaptive disentangled representation learning with the graph-based structure information. *Inf. Sci.*, 657(C), February 2024.
- [15] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, 2018.
- [16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.
- [17] Diederik P. Kingma, Danilo J. Rezende, Shakir Mohamed, and Max Welling. Semi-supervised Learning with Deep Generative Models. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.
- [18] Sean Kulinski and David I. Inouye. Towards explaining distribution shifts. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.
- [19] Yibo Li, Liangren Zhang, and Zhenming Liu. Multi-objective de novo drug design with conditional graph generative model. *Journal of Cheminformatics*, 10(1):33, 2018.

References V

- [20] Zinan Lin, Kiran Thekumparampil, Giulia Fanti, and Sewoong Oh. InfoGAN-CR and ModelCentrality: Self-supervised model training and selection for disentangling GANs. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6127–6139. PMLR, 13–18 Jul 2020.
- [21] Francesco Locatello, Stefan Bauer, Mario Lucic, Sylvain Gelly, Bernhard Schoelkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.
- [22] Francesco Locatello, Ben Poole, Gunnar Raetsch, Bernhard Schoelkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.
- [23] Pierre-Alexandre Mattei and Jes Frellsen. Leveraging the exact likelihood of deep latent variable models. *Advances in Neural Information Processing Systems*, 31, 2018.

References VI

- [24] Siddharth Narayanaswamy, Brooks Paige, Jan-Willem van de Meent, Alban Desmaison, Noah Goodman, Pushmeet Kohli, Frank Wood, and Philip Torr. Learning Disentangled Representations with Semi-Supervised Deep Generative Models. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [25] Weili Nie, Tero Karras, Animesh Garg, Shoubhik Debnath, Anjul Patney, Ankit Patel, and Animashree Anandkumar. Semi-supervised StyleGAN for disentanglement learning. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7360–7369. PMLR, 13–18 Jul 2020.
- [26] Gavin Perry, ET Rolls, and SM Stringer. Continuous transformation learning of translation invariant representations. *Experimental brain research*, 204(2):255–270, 2010.
- [27] Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly supervised disentanglement with guarantees. In *International Conference on Learning Representations (ICLR)*, 2020.

References VII

- [28] Junlin Yang, Nicha C Dvornek, Fan Zhang, Julius Chapiro, MingDe Lin, and James S Duncan. Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 255–263. Springer, 2019.
- [29] Linlin Yang and Angela Yao. Disentangling latent hands for image synthesis and pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [30] Fei Ye and Adrian G. Bors. Learning latent representations across multiple data domains using lifelong vaegan. In *European Conference on Computer Vision (ECCV)*, 2020.
- [31] Fei Ye and Adrian G. Bors. Lifelong teacher-student network learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6280–6296, October 2022.
- [32] Cheng Zhang, Kun Zhang, and Yingzhen Li. A causal view on robustness of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Overview

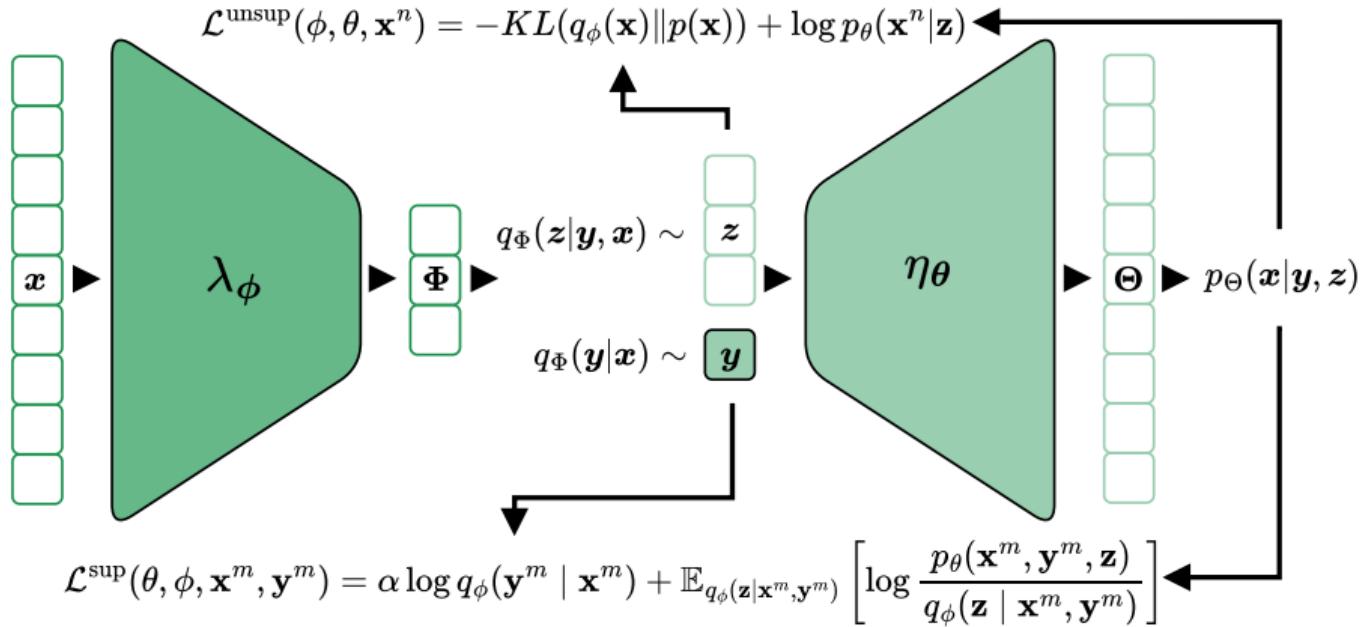


Figure: Semi-supervised disentanglement framework

Semi-Supervised Objective Formulation

The total objective combines unsupervised (x) and supervised (x, y) data, weighted by γ :

$$\mathcal{L}(\theta, \phi, \mathcal{D}, \mathcal{D}^{\text{sup}}) = \sum_{x^n \in \mathcal{D}}^N \mathcal{L}^{\text{unsup}}(\theta, \phi; x^n) + \gamma \sum_{(x^m, y^m) \in \mathcal{D}^{\text{sup}}}^M \mathcal{L}^{\text{sup}}(\theta, \phi; x^m, y^m)$$

The Supervised Term \mathcal{L}^{sup} : Defined to jointly maximize the generative likelihood and discriminative power:

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

Semi-Supervised Objective Formulation

In the supervised term, we cannot evaluate $q_\phi(z|x, y)$ directly:

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

We use that $q_\phi(z|x, y)$ factorizes to $\frac{q_\phi(y, z|x)}{q_\phi(y|x)}$ and get:

$$\mathcal{L}^{\text{sup}}(\theta, \phi; x, y) = (1 + \alpha) \log q_\phi(y|x) + \mathbb{E}_{q_\phi(z|x, y)} \left[\log \frac{p_\theta(x, y, z)}{q_\phi(y, z|x)} \right]$$

Now we approximate the expectation and $\log q_\phi(y|x)$ with importance sampling and get:

$$\widehat{\mathcal{L}}^{\text{sup}} = \sum_{s=1}^S \frac{w_s}{\sum_j w_j} \log \frac{p_\theta(x, y, z_s)}{q_\phi(y, z_s | x)} + (1 + \alpha) \log w_s$$

Approximation with Importance Sampling

Approximate the expectation with importance sampling:

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

Here we sample $z^s \sim q_\phi(z|x)$ from the unconditioned encoder with importance weights:

$$w^s := \frac{q_\phi(y,z^s|x)}{q_\phi(z^s|x)}, \quad Z = \frac{1}{S} \sum_{s=1}^S w^s$$

Using the same weights we approximate $\log q_\phi(y^m|x^m)$ with a Monte Carlo estimate of the lower bound:

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

Shu et al. [27] refined the definition of disentanglement into two distinct properties:

1. Consistency

- ▶ Degree to which representation is deterministic w.r.t. ground-truth factors
- ▶ If ground-truth factor (e.g., color) is fixed → latent code should be constant

2. Restrictiveness

- ▶ Degree to which single latent dimension encodes only one ground-truth factor
- ▶ Prevents single dimension from encoding both color and shape

Critique of Narayanaswamy et al.

False claim: Semi-supervision leads to disentangled results

Reality: Creates consistent representations, not necessarily restrictive

Caveat: Note that on real-world data, consistency and restrictiveness are often correlated

