On Disentanglement and Mutual Information in Semi-Supervised Variational Auto-Encoders

Elliott Gordon Rodríguez
Department of Statistics
Columbia University
eg2912@columbia.edu

Abstract

In the context of variational auto-encoders, learning disentangled latent variable representations remains a challenging problem. In this abstract, we consider the semi-supervised setting, in which the factors of variation are labelled for a small fraction of our samples. We examine how the quality of learned representations is affected by the dimension of the unsupervised component of the latent space. We also consider a variational lower bound for the mutual information between the data and the semi-supervised component of the latent space, and analyze its role in the context of disentangled representation learning.

1. Introduction

Many recent works have focused on improving the interpretability of latent variable representations in the variational auto-encoder (VAE) framework [18, 19]. Notably, the β-VAE optimizes a modified objective, where the KL regularization term in the *evidence lower bound* (ELBO) is upweighted in order to increase statistical independence in the latent space [10]. Other augmentations of the ELBO have been explored, similarly designed to encourage the desired properties in the posterior distribution of the latent variables [2, 16, 5, 7, 6]. In the semi-supervised setting, the labelled datapoints can be used to construct a supervised penalty term that is also added to the objective function [17, 28, 23]. In some – but not all – cases, this partial supervision can lead to disentangled generative models [20, 3, 14, 27, 22].

While unsupervised models have the benefit of requiring no labeled data, the identification of meaningful latent factors requires manual inspection of latent traversals for each model of interest. Given the rotational invariance of the priors that are commonly used in such models, this identification is sensitive to initialization, amongst other difficulties [24, 26]. On the other hand, semi-supervised VAEs offer the possibility to pre-specify latent components using the labelled datapoints [17, 28, 4, 30].

In this abstract, we evaluate the quality of the learned representations in the semi-supervised VAE, as the dimensionality of the unsupervised latent component is varied. We demonstrate empirically that, given sufficient capacity, the semi-supervised component of the latent space is ignored by the decoder. This phenomenon occurs regardless of whether the encoder model provides an accurate estimation of the semi-supervised latents. As a result, regulating the dimension of the latent space controls a tradeoff between disentangling and reconstruction quality.

In addition, we propose a novel modification of the ELBO, designed to maximize the mutual information between the semi-supervised latent variables and the decoder outputs. While the mutual information is intractable in the models of interest, we can construct a variational bound [1] that results in a differentiable objective. We also show that, for the VAE, this bound is equivalent to enforcing *cycle consistency* in the latent space, an idea with precedent in deep generative models [15, 33, 14, 27].

2. Semi-supervised Variational Autoencoder

Given a set of unlabelled samples, $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i\in\mathcal{U}}$, together with a subset of labelled pairs $\mathcal{D}^{sup} = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i\in\mathcal{S}}$, our goal is to learn a generative model of the form [17]:

(i) $(\mathbf{z}, \mathbf{y}) \sim p(\mathbf{z}, \mathbf{y})$,
(ii) $\mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{y})$.

Typically, we model $p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{y}) = \mathcal{N}(\mu_\theta(\mathbf{z}, \mathbf{y}), \Sigma_\theta(\mathbf{z}, \mathbf{y}))$, where μ_θ and Σ_θ denote deep networks parameterized by θ. In this setting, the marginal likelihood is intractable, as is the posterior. Therefore, the true posterior is approximated by a variational family $q_\phi(\mathbf{z}, \mathbf{y}|\mathbf{x})$, typically also parameterized by a neural network. The model parameters θ and ϕ can then be trained jointly by maximizing the evidence lower bound on the marginal likelihood. For the unlabelled

samples, the ELBO takes the standard form:

$$\mathcal{L}(\theta, \phi; \mathbf{x}^{(i)}) := \mathbb{E}_{q_\phi(\mathbf{z},\mathbf{y}|\mathbf{x}^{(i)})}\left[\log \frac{p_\theta(\mathbf{x}^{(i)}, \mathbf{z}, \mathbf{y})}{q_\phi(\mathbf{z}, \mathbf{y}|\mathbf{x}^{(i)})}\right] \leq \log p_\theta(\mathbf{x}^{(i)}). \quad (1)$$

As for the supervised samples, where $\mathbf{y}^{(i)}$ is observed, note that:

$$\log p_\theta(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^{(i)},\mathbf{y}^{(i)})}\left[\log \frac{p_\theta(\mathbf{x}^{(i)}, \mathbf{z}, \mathbf{y}^{(i)})}{q_\phi(\mathbf{z}|\mathbf{x}^{(i)}, \mathbf{y}^{(i)})}\right] - \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}^{(i)}, \mathbf{y}^{(i)})||p_\theta(\mathbf{z}|\mathbf{x}^{(i)}, \mathbf{y}^{(i)})). \quad (2)$$

Thus, using the non-negativity of the KL term, a similar lower bound can be constructed:

$$E_{q_\phi(\mathbf{z}|\mathbf{x}^{(i)},\mathbf{y}^{(i)})}\left[\log \frac{p_\theta(\mathbf{x}^{(i)}, \mathbf{z}, \mathbf{y}^{(i)})}{q_\phi(\mathbf{z}|\mathbf{x}^{(i)}, \mathbf{y}^{(i)})}\right] \leq \log p_\theta(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}). \quad (3)$$

The label information is further incorporated in a supervised loss term, $\log q_\phi(\mathbf{y}^{(i)}|\mathbf{x}^{(i)})$, to encourage the encoder to learn a good mapping $\mathbf{x} \mapsto \mathbf{y}$:

$$\mathcal{L}^{sup}(\theta, \phi; \mathbf{x}^{(i)}, \mathbf{y}^{(i)}) := E_{q_\phi(\mathbf{z}|\mathbf{x}^{(i)},\mathbf{y}^{(i)})}\left[\log \frac{p_\theta(\mathbf{x}^{(i)}, \mathbf{z}, \mathbf{y}^{(i)})}{q_\phi(\mathbf{z}|\mathbf{x}^{(i)}, \mathbf{y}^{(i)})}\right] + \alpha \cdot \log q_\phi(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}). \quad (4)$$

An aggregated objective over all data can then be constructed:

$$\mathcal{L}(\theta, \phi; \mathcal{D}, \mathcal{D}^{sup}) = \sum_{i \in \mathcal{U}} \mathcal{L}(\theta, \phi; \mathbf{x}^{(i)}) + \sum_{i \in \mathcal{S}} \mathcal{L}^{sup}(\theta, \phi; \mathbf{x}^{(i)}, \mathbf{y}^{(i)}). \quad (5)$$

3. Disentanglement and Mutual Information

Semi-supervised VAEs can struggle to disentangle a generative factor of interest, even when (partial) label information is available for such a factor [28]. As we will demonstrate empirically, the key difficulty lies in the mapping $\mathbf{y} \mapsto \mathbf{x}$. Intuitively, the autoencoder $\mathbf{x} \mapsto (\mathbf{z}, \mathbf{y}) \mapsto \mathbf{x}$ can separately learn a regression mapping $\mathbf{x} \mapsto \mathbf{y}$ and a reconstruction mapping $\mathbf{x} \mapsto \mathbf{z} \mapsto \mathbf{x}$. As a result, the decoder will ignore the latent variable $\mathbf{y}$, failing to disentangle the generative factors of interest. In other words, a disentangled factor $\mathbf{y}^{(i)}$ need not provide an ideal input for the decoder to obtain a good reconstruction of $\mathbf{x}^{(i)}$. If so, the decoder will tend to focus exclusively on its first input, $\mathbf{z}$.

In order to address this shortcoming, we propose augmenting the semi-supervised objective (Eq. 5) with a mutual information term between $\mathbf{x}$ and $\mathbf{y}$ (but not $\mathbf{z}$). The mutual information is taken under the generative model, in other words, between the semi-supervised component of the decoder input, and the decoder output. Intuitively, this term encourages the information available in $\mathbf{y}$, which contains the disentangled latent variables of interest, to flow through the decoder. Previous works have evaluated information-theoretic criteria in variational autoencoders [32, 8, 11, 31, 29]; here we apply a similar ideas specifically to the semi-supervised component of the latent space.

Let us denote the mutual information under the generative model as:

$$I(\mathbf{x}, \mathbf{y}) = H(\mathbf{y}) - H(\mathbf{y}|\mathbf{x}), \quad (6)$$

where $H(\mathbf{y}) = -\mathbb{E}_{p(\mathbf{y})}[\log p(\mathbf{y})]$ and $H(\mathbf{y}|\mathbf{x}) = -\mathbb{E}_{p_\theta(\mathbf{y},\mathbf{x})}[\log p_\theta(\mathbf{y}|\mathbf{x})]$ denote the entropy of $\mathbf{y}$ and the conditional entropy of $\mathbf{y}|\mathbf{x}$, respectively. While the first term $H(\mathbf{y})$ is a constant and can be ignored for the purposes of optimization, the second term $H(\mathbf{y}|\mathbf{x})$ involves the intractable posterior and cannot be computed directly. However, we can construct a variational lower bound [1], again using the non-negativity of the KL divergence:

$$\begin{aligned} -H(\mathbf{y}|\mathbf{x}) &= \mathbb{E}_{p_\theta(\mathbf{y},\mathbf{x})}[\log p_\theta(\mathbf{y}|\mathbf{x})] \\ &= \mathbb{E}_{p_\theta(\mathbf{y},\mathbf{x})}[\log q_\phi(\mathbf{y}|\mathbf{x})] \\ &\quad + \mathbb{E}_{p_\theta(\mathbf{y},\mathbf{x})}\left[\frac{\log p_\theta(\mathbf{y}|\mathbf{x})}{\log q_\phi(\mathbf{y}|\mathbf{x})}\right] \\ &= \mathbb{E}_{p_\theta(\mathbf{y},\mathbf{x})}[\log q_\phi(\mathbf{y}|\mathbf{x})] \\ &\quad + \mathbb{E}_{p_\theta(\mathbf{x})}[KL(p_\theta(\mathbf{y}|\mathbf{x})||q_\phi(\mathbf{y}|\mathbf{x})] \\ &\geq \mathbb{E}_{p_\theta(\mathbf{y},\mathbf{x})}[\log q_\phi(\mathbf{y}|\mathbf{x})] \\ &=: \tilde{I}(\theta, \phi). \end{aligned} \quad (7)$$

Importantly, this bound no longer depends on the intractable posterior, allowing for efficient gradient optimization via the reprameterization trick [18]. Namely, we can construct a differentiable Monte Carlo estimate as follows:

1. Draw samples of $(\mathbf{z}, \mathbf{y})$ from the prior.
2. Feed them through the decoder network, and add the appropriate noise distribution (as per the reparametrization trick) to generate synthetic samples of $\mathbf{x}$.
3. Feed the synthetic samples back through the encoder network to obtain "reconstructed latents" $(\hat{\mathbf{z}}, \hat{\mathbf{y}})$.
4. Compute the "latent error" between the original $\mathbf{y}$ and the reconstructed $\hat{\mathbf{y}}$, as measured by the log-probability of q_ϕ (in the normal case, $-\|\mathbf{y} - \hat{\mathbf{y}}\|^2$). Note that $\hat{\mathbf{z}}$ is not used for estimating $\tilde{I}(\theta, \phi)$.

This procedure admits a simple interpretation: for the mutual information to be high, the values of $\mathbf{y}$ should affect the decoder output in such a way that the encoder can, in turn, map back to the values of $\mathbf{y}$ that generated such an output.

Model	Identity error	Lighting error	Log-likelihood	Reconstruction RMSE	Mutual Information
Baseline	3.5% ($\pm$3.4)	17.6% ($\pm$1.8)	214.4 ($\pm$3.7)	0.70% ($\pm$0.03)	-0.15 ($\pm$0.02)
$\dim(\mathbf{z})$=10	3.3% ($\pm$2.0)	17.0% ($\pm$3.7)	211.1 ($\pm$4.7)	0.74% ($\pm$0.05)	-0.16 ($\pm$0.02)
$\dim(\mathbf{z})$=2	3.2% ($\pm$2.8)	11.9% ($\pm$8.7)	184.2 ($\pm$22.8)	1.05% ($\pm$0.29)	-0.08 ($\pm$0.06)
γ=0.1	4.3% ($\pm$3.1)	8.3% ($\pm$1.1)	202.0 ($\pm$6.4)	0.79% ($\pm$0.05)	-0.04 ($\pm$0.01)
γ=1.0	6.5% ($\pm$3.5)	7.2% ($\pm$1.7)	199.2 ($\pm$14.4)	0.82% ($\pm$0.13)	-0.03 ($\pm$0.02)
γ=10.0	8.2% ($\pm$3.9)	7.5% ($\pm$1.9)	163.3 ($\pm$18.6)	5.82% ($\pm$0.58)	-0.02 ($\pm$0.03)
Fully-sup	1.9% ($\pm$1.5)	3.1% ($\pm$3.8)	222.7 ($\pm$4.1)	0.59% ($\pm$0.03)	-0.05 ($\pm$0.04)

Table 1. Evaluation metrics on held-out data. Estimation errors shown in parenthesis correspond to two standard errors over 16 random initializations of the model. The descriptor in the first column shows where each model differs from the baseline model. The last column shows a Monte Carlo estimate of the mutual information lower bound (Eq. 7), a quantity that depends only on the trained model and not the held-out data. Note that the mutual information is applied to the lighting component only (since the baseline model was able to disentangle the identity component).

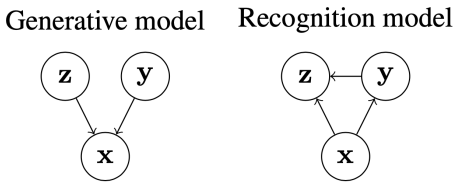

Figure 1. Conditional dependence structure for our baseline semi-supervised VAE [28]. $\mathbf{x}$ represents an image sample, $\mathbf{y}$ encodes the identity of the subject and the lighting of the image, and $\mathbf{z}$ encodes all other generative factors.

In other words, our mutual information criterion reduces to a measure of *cycle consistency* in the latent space, an idea in the spirit of [15, 33, 14, 27], but applied specifically to the semi-supervised component $\mathbf{y}$. Adding the mutual information lower bound (Eq. 7) to our objective function (Eq. 5) gives rise to a new optimization problem:

$$\min_{\theta,\phi}\{\mathcal{L}(\theta,\phi;\mathcal{D},\mathcal{D}^{sup})+\gamma\tilde{I}(\theta,\phi)\}, \tag{8}$$

where γ is a hyperparameter that controls the relative strength of the mutual information term.

4. Experiments

Our experimental setup uses the Extended Yale Face Database B [9, 21], as processed by [12]. This dataset contains a total of $\sim$1,700 images of 38 subjects under 45 illumination conditions. All samples are labelled with the identity of the subject and the lighting angle of incidence, allowing us to fit both semi-supervised and fully-supervised models.

Our baseline model is a semi-supervised VAE composed of a 4-layer encoder, a 4-layer decoder, and an additional layer mapping $\mathbf{y}$ to $\mathbf{z}$,[1] with $\dim(\mathbf{y}) = 39$ and $\dim(\mathbf{z}) = 20$ (Figure 1) [28]. The unsupervised latent components $\mathbf{z}$ follow independent standard normal prior distributions. The semi-supervised latent $\mathbf{y}$ is composed by a 38-level categorical variable modeling the identity of the subject,[2] and an additional scalar variable modeling the lighting of the image, i.e., the angle of incidence. 15% of the labels were made available during training.

Our evaluation criteria include:

(a) Reconstruction quality, measured both qualitatively and quantitatively, by log-likehood and reconstruction RMSE.
(b) Classification accuracy for the identity component of $\mathbf{y}$.
(c) Regression error for the lighting component of $\mathbf{y}$.
(d) Disentanglement, measured qualitatively as well as through the mutual information lower bound of Equation 7.

Our results, summarized in Table 1, compare the baseline model against:

(i) A fully-supervised objective, where 100% of the labels are made available during training (Figure 2).
(ii) Models with reduced latent dimension (Figure 3).
(iii) An objective with the mutual information term, at varying strengths γ (Figure 4).

5. Discussion

Figure 3 illustrates that regulating the dimensionality of the unsupervised latent component is highly effective for achieving disentangled representations. Intuitively, a low-dimensional $\mathbf{z}$ corresponds to a narrow "information bottleneck", encouraging the decoder to draw more heavily on the information encoded in $\mathbf{y}$. In fact, sufficiently reducing $\dim(\mathbf{z})$ resulted in a disentangled representation for both

[1] We found empirically that such a layer had minimal impact on results.

[2] During optimization, categorical latent variables are relaxed via the Gumbel-Softmax distribution [13, 25].

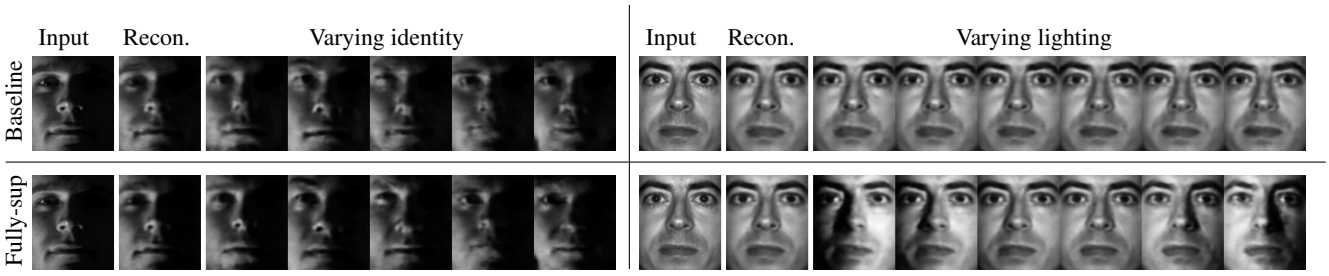

Figure 2. For a fixed input image, we show the model reconstruction, as well as latent traversals obtained by varying the identity and lighting components of $\mathbf{y}$. The baseline model (semi-supervised) fails to disentangle the lighting factor, which becomes possible under full supervision (bottom right).

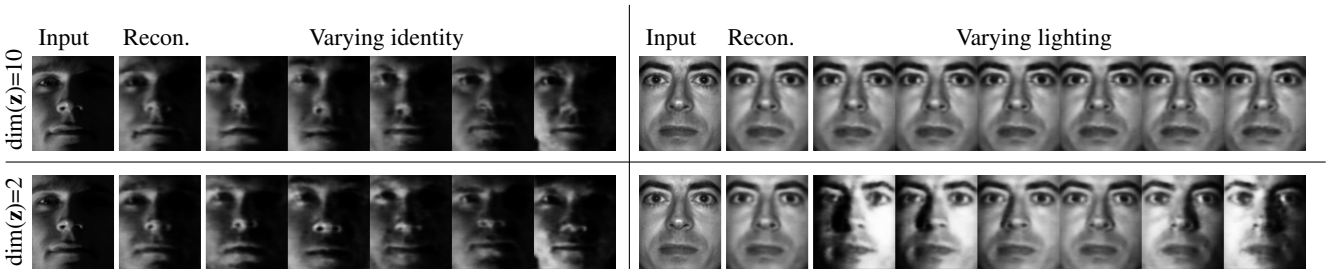

Figure 3. Similar to Figure 2, for two semi-supervised models of reduced latent dimension. The semi-supervised model with just 2 dimensions for $\mathbf{z}$ (bottom row) achieves a disentangled representation for both identity and lighting, of comparable quality to the fully-supervised model (Figure 2).

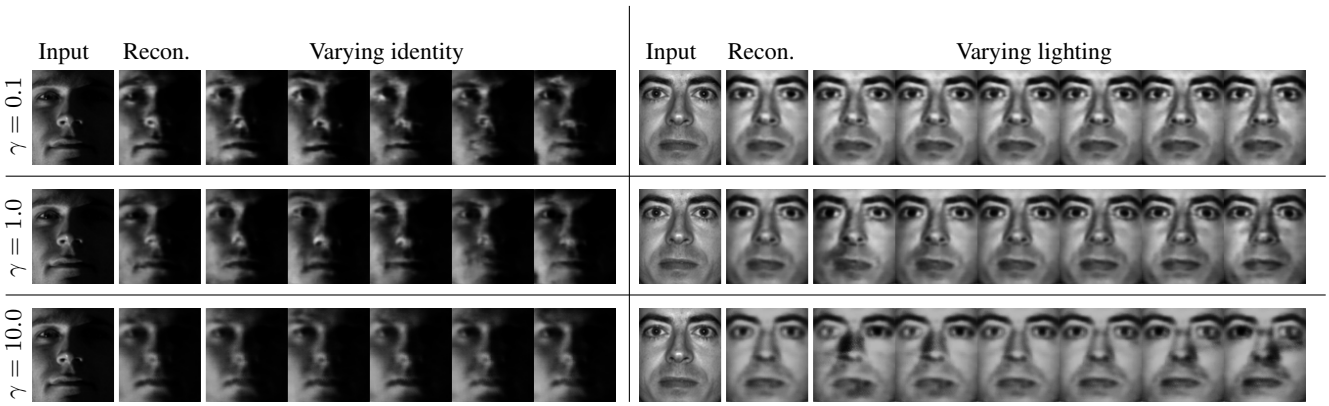

Figure 4. Similar to Figure 2, for three semi-supervised models with varying strength γ applied to the mutual information term in the objective function. Note the mutual information is applied to the lighting component only. As its strength increases, the reconstruction quality degrades with little improvement in disentanglement.

identity and lighting, of similar quality to the fully supervised model (compare Figures 2 and 3). However, there is some tradeoff between disentanglement and reconstruction quality, since a lower $\dim(\mathbf{z})$ implies a less flexible model, in this case leading to a small increase in reconstruction error (top 3 rows of Table 1). It is also worth remarking that a reduction in $\dim(\mathbf{z})$ improves the out-of-sample accuracy of the semi-supervised mapping $\mathbf{x} \mapsto \mathbf{y}$, even though its network architecture remains unchanged (of course, its gradients will change due to shared weights and biases). This suggests that the autoencoder $\mathbf{x} \mapsto (\mathbf{z}, \mathbf{y}) \mapsto \mathbf{x}$ can act as a regularizer on the classifier $\mathbf{x} \mapsto \mathbf{y}$. On the other hand, the mutual information term (Eq. 7) only provides small improvements toward disentangling the latent space, at the expense of a significant decrease in reconstruction quality (Figure 4). This behavior indicates a similar tradeoff between reconstruction quality and cycle consistency in the latent space. Taken together, these results indicate that learning disentangled representations requires a fine balancing act between model architecture, objective function, and partial supervision.

6. Acknowledgements

We thank John Cunningham for inspiration and support.

References

[1] David Barber Felix Agakov. The im algorithm: a variational approach to information maximization. *Advances in Neural Information Processing Systems*, 16:201, 2004. 1, 2

[2] Alexander A Alemi, Ben Poole, Ian Fischer, Joshua V Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken elbo. *arXiv preprint arXiv:1711.00464*, 2017. 1

[3] Ershad Banijamali, Amir-Hossein Karimi, Alexander Wong, and Ali Ghodsi. Jade: Joint autoencoders for disentanglement. *arXiv preprint arXiv:1711.09163*, 2017. 1

[4] Junxiang Chen and Kayhan Batmanghelich. Weakly supervised disentanglement by pairwise similarities. *arXiv preprint arXiv:1906.01044*, 2019. 1

[5] Tian Qi Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 2615–2625, 2018. 1

[6] Jan Chorowski, Ron J Weiss, Samy Bengio, and Aäron van den Oord. Unsupervised speech representation learning using wavenet autoencoders. *IEEE/ACM transactions on audio, speech, and language processing*, 27(12):2041–2053, 2019. 1

[7] Babak Esmaeili, Hao Wu, Sarthak Jain, Alican Bozkurt, N Siddharth, Brooks Paige, Dana H Brooks, Jennifer Dy, and Jan-Willem van de Meent. Structured disentangled representations. *stat*, 1050:12, 2018. 1

[8] Shuyang Gao, Rob Brekelmans, Greg Ver Steeg, and Aram Galstyan. Auto-encoding total correlation explanation. *arXiv preprint arXiv:1802.05822*, 2018. 2

[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. 3

[10] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. 1

[11] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018. 2

[12] Varun Jampani, SM Ali Eslami, Daniel Tarlow, Pushmeet Kohli, and John Winn. Consensus message passing for layered graphical models. In *Artificial Intelligence and Statistics*, pages 425–433, 2015. 3

[13] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 3

[14] Ananya Harsh Jha, Saket Anand, Maneesh Singh, and VSR Veeravasarapu. Disentangling factors of variation with cycle-consistent variational auto-encoders. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 805–820, 2018. 1, 3

[15] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas. Forward-backward error: Automatic detection of tracking failures. In *2010 20th International Conference on Pattern Recognition*, pages 2756–2759. IEEE, 2010. 1, 3

[16] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018. 1

[17] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589, 2014. 1

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1, 2

[19] Diederik P Kingma and Max Welling. An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*, 2019. 1

[20] Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *Advances in neural information processing systems*, pages 2539–2547, 2015. 1

[21] Kuang-Chih Lee, Jeffrey Ho, and David J Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (5):684–698, 2005. 3

[22] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, and Mingwu Ren. Gait recognition via semi-supervised disentangled representation learning to identity and covariate features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13309–13319, 2020. 1

[23] Yang Li, Quan Pan, Suhang Wang, Haiyun Peng, Tao Yang, and Erik Cambria. Disentangled variational auto-encoder for semi-supervised learning. *Information Sciences*, 482:73–85, 2019. 1

[24] Francesco Locatello, Stefan Bauer, Mario Lucic, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359*, 2018. 1

[25] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016. 3

[26] Emile Mathieu, Tom Rainforth, N Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders. In *International Conference on Machine Learning*, pages 4402–4412. PMLR, 2019. 1

[27] Tasnim Mohiuddin and Shafiq Joty. Revisiting adversarial autoencoder for unsupervised word translation with cycle consistency and improved training. *arXiv preprint arXiv:1904.04116*, 2019. 1, 3

[28] Siddharth Narayanaswamy, T 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*, pages 5925–5935, 2017. 1, 2, 3

[29] Ali Lotfi Rezaabad and Sriram Vishwanath. Learning representations by maximizing mutual information in variational autoencoders. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2729–2734. IEEE, 2020. 2

[30] Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly supervised disentanglement with guarantees. *arXiv preprint arXiv:1910.09772*, 2019. 1

[31] Shujian Yu and Jose C Principe. Understanding autoencoders with information theoretic concepts. *Neural Networks*, 117:104–123, 2019. 2

[32] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Information maximizing variational autoencoders. *arXiv preprint arXiv:1706.02262*, 2017. 2

[33] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1, 3