VARIATIONAL INFERENCE OF DISENTANGLED LATENT CONCEPTS FROM UNLABELED OBSERVATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Disentangled representations, where the higher level data generative factors are reflected in disjoint latent dimensions, offer several benefits such as ease of deriving invariant representations, transferability to other tasks, interpretability, etc. We consider the problem of unsupervised learning of disentangled representations from large pool of unlabeled observations, and propose a variational inference based approach to infer disentangled latent factors. We introduce a regularizer on the expectation of the approximate posterior over observed data that encourages the disentanglement. We evaluate the proposed approach using several quantitative metrics and empirically observe significant gains over existing methods in terms of both disentanglement and data likelihood (reconstruction quality).

1 INTRODUCTION

Feature representations of the observed raw data play a crucial role in the success of machine learning algorithms. Effective representations should be able to capture the underlying (abstract or high-level) latent generative factors that are relevant for the end task while ignoring the inconsequential or nuisance factors. *Disentangled* feature representations have the property that the generative factors are revealed in disjoint subsets of the feature dimensions, such that a change in a single generative factor causes a highly sparse change in the representation. Disentangled representations offer several advantages – *(i) Invariance:* it is easier to derive representations that are invariant to nuisance factors by simply marginalizing over the corresponding dimensions, *(ii) Transferability:* they are arguably more suitable for transfer learning as most of the key underlying generative factors appear segregated along feature dimensions, *(iii) Interpretability:* a human expert may be able to assign meanings to the dimensions, *(iv) Conditioning and intervention:* they allow for interpretable conditioning and/or intervention over a subset of the latents and observe the effects on other nodes in the graph. Indeed, the importance of learning disentangled representations has been argued in several recent works (Bengio et al., 2013; Lake et al., 2016; Ridgeway, 2016).

Recognizing the significance of disentangled representations, several attempts have been made in this direction in the past (Ridgeway, 2016). Much of the earlier work assumes some sort of supervision in terms of: (i) partial or full access to the generative factors per instance (Reed et al., 2014; Yang et al., 2015; Kulkarni et al., 2015; Karaletsos et al., 2015), (ii) knowledge about the nature of generative factors (e.g, translation, rotation, etc.) (Hinton et al., 2011; Cohen & Welling, 2014), (iii) knowledge about the changes in the generative factors across observations (e.g., sparse changes in consecutive frames of a Video) (Goroshin et al., 2015; Whitney et al., 2016; Fraccaro et al., 2017; Denton & Birodkar, 2017; Hsu et al., 2017), (iv) knowledge of a complementary signal to infer representations that are conditionally independent of it[1] Cheung et al. (2014); Mathieu et al. (2016); Siddharth et al. (2017), However, in most real scenarios, we only have access to raw observations without any supervision about the generative factors. It is a challenging problem and many of the earlier attempts have not been able to scale well for realistic settings (Cohen & Welling, 2015) (see also, Higgins et al. (2017)).

Recently, Chen et al. (2016) proposed an approach to learn a generative model with disentangled factors based on Generative Adversarial Networks (GAN) (Goodfellow et al., 2014), however

[1]The representation itself can still be entangled in rest of the generative factors.

implicit generative models like GANs lack an effective inference mechanism[2], which hinders its applicability to the problem of learning disentangled representations. More recently, Higgins et al. (2017) proposed an approach based on Variational AutoEncoder (VAE) Kingma & Welling (2013) for inferring disentangled factors. The inferred latents using their method (termed as β-VAE) are empirically shown to have better disentangling properties, however the method deviates from the basic principles of variational inference, creating increased tension between observed data likelihood and disentanglement. This in turn leads to poor quality of generated samples as observed in (Higgins et al., 2017).

In this work, we propose a principled approach for inference of disentangled latent factors based on the popular and scalable framework of amortized variational inference (Kingma & Welling, 2013; Stuhlmüller et al., 2013; Gershman & Goodman, 2014; Rezende et al., 2014) powered by stochastic optimization (Hoffman et al., 2013; Kingma & Welling, 2013; Rezende et al., 2014). Disentanglement is encouraged by introducing a regularizer over the induced *inferred prior*. Unlike β-VAE (Higgins et al., 2017), our approach does not introduce any extra conflict between disentanglement of the latents and the observed data likelihood, which is reflected in the overall quality of the generated samples that matches the VAE and is much better than β-VAE. This does *not* come at the cost of higher entanglement and our approach also outperforms β-VAE in disentangling the latents as measured by various quantitative metrics.

2 FORMULATION

We start with a generative model of the observed data that first samples a latent variable $\mathbf{z} \sim p(\mathbf{z})$, and an observation is generated by sampling from $p_\theta(\mathbf{x}|\mathbf{z})$. The joint density of latents and observations is denoted as $p_\theta(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p_\theta(\mathbf{x}|\mathbf{z})$. The problem of inference is to compute the posterior of the latents conditioned on the observations, i.e., $p(\mathbf{z}|\mathbf{x})$. We assume that we are given a finite set of samples (observations) from the true data distribution $p(\mathbf{x})$. In most practical scenarios involving high dimensional and complex data, this computation is intractable and calls for approximate inference. Variational inference takes an optimization based approach to this, positing a family $\mathcal{D}$ of approximate densities over the latents and reducing the approximate inference problem to finding a member density that minimizes the Kullback-Leibler divergence to the true posterior, i.e., $q^*_{\mathbf{x}} = \min_{q \in \mathcal{D}} \mathrm{KL}(q(\mathbf{z}) \| p(\mathbf{z}|\mathbf{x}))$ (Blei et al., 2017). The idea of amortized inference (Kingma & Welling, 2013; Stuhlmüller et al., 2013; Gershman & Goodman, 2014; Rezende et al., 2014) is to explicitly share information across inferences made for each observation. One successful way of achieving this for variational inference is to have a so-called *recognition model*, parameterized by ϕ, that encodes an inverse map from the observations to the approximate posteriors (also referred as variational autoencoder or VAE) (Kingma & Welling, 2013; Rezende et al., 2014). The recognition model parameters are learned by optimizing the problem $\min_\phi \mathbb{E}_{\mathbf{x}} \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z}|\mathbf{x}))$. This can be shown as equivalent to maximizing what is termed as evidence lower bound (ELBO):

$$\arg\min_{\theta,\phi} \mathbb{E}_{\mathbf{x}} \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z}|\mathbf{x})) = \arg\max_{\theta,\phi} \mathbb{E}_{\mathbf{x}} \left[\mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} \left[\log p_\theta(\mathbf{x}|\mathbf{z}) \right] - \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})) \right] \quad (1)$$

The ELBO (the objective at the right side of Eq. 1) lower bounds the log-likelihood of observed data, and the gap vanishes at the global optimum. Often, the density forms of $p(\mathbf{z})$ and $q_\phi(\mathbf{z}|\mathbf{x})$ are chosen such that their KL-divergence can be written analytically in a closed-form expression (e.g., $p(\mathbf{z})$ is $N(0, I)$ and $q_\phi(\mathbf{z}|\mathbf{x})$ is $N(\mu_\phi(\mathbf{x}), \Sigma_\phi(\mathbf{x}))$) (Kingma & Welling, 2013). In such cases, the ELBO can be efficiently optimized (to a stationary point) using stochastic first order methods where both expectations are estimated using mini-batches. Further, in cases when $q_\phi(\cdot)$ can be written as a continuous transformation of a fixed base distribution (e.g., the standard normal distribution), a low variance estimate of the gradient over ϕ can be obtained by coordinate transformation (also referred as reparametrization) (Fu, 2006; Kingma & Welling, 2013; Rezende et al., 2014).

2.1 GENERATIVE STORY: DISENTANGLED PRIOR

Most VAE based generative models for real datasets (e.g., text, images, etc.) already work with a relatively simple and disentangled prior $p(\mathbf{z})$ having no interaction among the latent dimensions (e.g.,

[2]There have been a few recent attempts in this direction for visual data (Dumoulin et al., 2016; Donahue et al., 2016; Kumar et al., 2017) but often the reconstructed samples are semantically quite far from the input samples, sometimes even changing in the object classes.

the standard Gaussian $N(0, I)$) (Bowman et al., 2015; Miao et al., 2016; Hou et al., 2017; Zhao et al., 2017). The complexity of the observed data is absorbed in the conditional distribution $p_\theta(\mathbf{x}|\mathbf{z})$ which encodes the interactions among the latents. Hence, as far as the generative modeling is concerned, disentangled prior sets us in the right direction.

2.2 INFERRING DISENTANGLED LATENTS

Although the generative model starts with a disentangled prior, our main objective is to *infer* disentangled latents which are conducive for various goals mentioned in Sec. 1 (e.g., invariance, transferability, interpretability). To this end, we consider the density over the inferred latents induced by the approximate posterior inference mechanism,

$$q_\phi(\mathbf{z}) = \int q_\phi(\mathbf{z}|\mathbf{x})p(\mathbf{x})d\mathbf{x}, \tag{2}$$

which we will subsequently refer to as the *inferred prior*. For inferring disentangled factors, this should be factorizable along the dimensions, i.e., $q_\phi(\mathbf{z}) = \prod_i q_i(z_i)$, or equivalently $q_{i|j}(z_i|z_j) = q_i(z_i)$, $\forall i, j$. This can be achieved by minimizing a suitable distance between the inferred prior $q_\phi(\mathbf{z})$ and the disentangled generative prior $p(\mathbf{z})$. If we take KL-divergence as our choice of distance, by relying on its pairwise convexity (i.e., $\mathrm{KL}(\lambda p_1 + (1-\lambda)p_2 \| \lambda q_1 + (1-\lambda)q_2) \leq \lambda \mathrm{KL}(p_1\|q_1) + (1-\lambda)\mathrm{KL}(p_2\|q_2)$) (Van Erven & Harremos, 2014), we can show that this distance is upper bounded by the objective of the variational inference (the ELBO in Eq. (1)):

$$\mathrm{KL}(q_\phi(\mathbf{z})\|p(\mathbf{z})) = \mathrm{KL}(\mathbb{E}_\mathbf{x} q_\phi(\mathbf{z}|\mathbf{x})\|\mathbb{E}_\mathbf{x} p(\mathbf{z}|\mathbf{x})) \leq \mathbb{E}_\mathbf{x}\, \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}|\mathbf{x})). \tag{3}$$

Hence, variational posterior inference of latent variables with disentangled prior naturally encourages inferring factors that are disentangled. We think this is the reason that the original VAE (Eq. (1) has also been observed to exhibit some disentangling behavior on simple datasets such as MNIST (Kingma & Welling, 2013). However, this behavior does not carry over to more complex datasets (Aubry et al., 2014; Liu et al., 2015; Higgins et al., 2017), unless extra supervision on the generative factors is provided (Kulkarni et al., 2015; Karaletsos et al., 2015). This can be due to non-convexity of the ELBO objective which prevents us from achieving the global minimum of $\mathbb{E}_\mathbf{x}\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}|\mathbf{x})) = 0$ (which would imply $\mathrm{KL}(q_\phi(\mathbf{z})\|p(\mathbf{z})) = 0$). In other words, maximizing the ELBO (Eq. (1)) might also result in reducing the value of $\mathrm{KL}(q(\mathbf{z})\|p(\mathbf{z}))$, however, due to the non-convexity of the loss surface of the ELBO, the gap between $\mathrm{KL}(q_\phi(\mathbf{z})\|p(\mathbf{z}))$ and $\mathbb{E}_\mathbf{x}\, \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}|\mathbf{x}))$ could be large at the stationary point of convergence. Hence, minimizing $\mathrm{KL}(q_\phi(\mathbf{z})\|p(\mathbf{z}))$ explicitly will give us better control on the disentanglement. This motivates us to add $\mathrm{KL}(q_\phi(\mathbf{z})\|p(\mathbf{z}))$ as part of the objective to encourage disentanglement during inference, i.e.,

$$\max_{\theta,\phi} \mathbb{E}_\mathbf{x}\left[\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))\right] - \lambda\, \mathrm{KL}(q_\phi(\mathbf{z})\|p(\mathbf{z})), \tag{4}$$

where λ controls its contribution to the overall objective.

Optimizing (4) directly is not tractable due to the presence of $\mathrm{KL}(q_\phi(\mathbf{z})\|p(\mathbf{z}))$ which does not have a closed-form expression. One possibility is use the variational formulation of the KL-divergence (Nguyen et al., 2010; Nowozin et al., 2016) that needs only samples from $q_\phi(\mathbf{z})$ and $p(\mathbf{z})$ to estimate a lower bound to $\mathrm{KL}(q_\phi(\mathbf{z})\|p(\mathbf{z}))$. However, this would involve optimizing for a third set of parameters ψ for the KL-divergence estimator, and would also change the optimization to a saddle-point (min-max) problem which has its own challenges (e.g., gradient vanishing as encountered in training generative adversarial networks with KL or JS divergences (Goodfellow et al., 2014; Arjovsky & Bottou, 2017)). Replacing $\mathrm{KL}(q_\phi(\mathbf{z})\|p(\mathbf{z}))$ with another suitable distance between $q_\phi(\mathbf{z})$ and $p(\mathbf{z})$ (e.g., integral probability metrics like Wasserstein distance (Sriperumbudur et al., 2009)) might alleviate some of these issues (Arjovsky et al., 2017) but will still involve complicating the optimization to a saddle point problem in three set of parameters[3]. It should also be noted that using these variational forms of the distances will still leave us with an approximation to the actual distance.

We adopt a simpler alternative of matching the moments of the two distributions. In particular, we match the covariance of the two distributions which will amount to decorrelating the dimensions of

[3]Nonparametric distances like maximum mean discrepancy (MMD) with a characteristic kernel (Gretton et al., 2012) is also an option, however it has its own challenges when combined with stochastic optimization (Dziugaite et al., 2015; Li et al., 2015).

$\mathbf{z} \sim q_\phi(\mathbf{z})$ if $p(\mathbf{z})$ is $N(0, I)$. Let us denote $\text{Cov}_{q(\mathbf{z})}(\mathbf{z}) := \mathbb{E}_{q(\mathbf{z})}\left[(\mathbf{z} - \mathbb{E}_{q(\mathbf{z})}(\mathbf{z}))(\mathbf{z} - \mathbb{E}_{q(\mathbf{z})}(\mathbf{z}))^\top\right]$. By the law of total covariance, the covariance of $\mathbf{z} \sim q_\phi(\mathbf{z})$ is given by cite a book here

$$\text{Cov}_{q_\phi(\mathbf{z})}(\mathbf{z}) = \mathbb{E}_{p(\mathbf{x})}\text{Cov}_{q_\phi(\mathbf{z}|\mathbf{x})}(\mathbf{z}) + \text{Cov}_{p(\mathbf{x})}\left(\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}(\mathbf{z})\right), \tag{5}$$

where $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}(\mathbf{z})$ and $\text{Cov}_{q_\phi(\mathbf{z}|\mathbf{x})}(\mathbf{z})$ are random variables that are functions of the random variable $\mathbf{x}$ ($\mathbf{z}$ is marginalized over). Most existing work on the VAE models uses $q_\phi(\mathbf{z}|\mathbf{x})$ having the form $N(\boldsymbol{\mu}_\phi(\mathbf{x}), \boldsymbol{\Sigma}_\phi(\mathbf{x}))$, where $\boldsymbol{\mu}_\phi(\mathbf{x})$ and $\boldsymbol{\Sigma}_\phi(\mathbf{x})$ are the outputs of a deep neural net. In this case Eq. (5) reduces to $\text{Cov}_{q_\phi(\mathbf{z})}(\mathbf{z}) = \mathbb{E}_{p(\mathbf{x})}\boldsymbol{\Sigma}_\phi(\mathbf{x}) + \text{Cov}_{p(\mathbf{x})}(\boldsymbol{\mu}_\phi(\mathbf{x}))$, which we want to be close to an identity matrix. For simplicity, we choose entry-wise squared ℓ_2-norm as the measure of proximity. However, as the entanglement is mainly reflected in the off-diagonal entries of this matrix, we opt for two separate hyperparameters controlling the relative importance of the loss on the diagonal and off-diagonal entries. This gives rise to the following optimization problem for inference:

$$\max_{\theta,\phi} \text{ELBO}(\theta, \phi) - \lambda_{od} \sum_{i \neq j} \left[\text{Cov}_{p(\mathbf{x})}(\boldsymbol{\mu}_\phi(\mathbf{x}))\right]_{ij}^2 - \lambda_d \sum_i \left(\left[\text{Cov}_{p(\mathbf{x})}(\boldsymbol{\mu}_\phi(\mathbf{x}))\right]_{ii} - 1\right)^2. \tag{6}$$

The regularization terms involving $\text{Cov}_{p(\mathbf{x})}(\boldsymbol{\mu}_\phi(\mathbf{x}))$ in the above objective (6) can be efficiently optimized using SGD. We maintain a running estimate of $\text{Cov}_{p(\mathbf{x})}(\boldsymbol{\mu}_\phi(\mathbf{x}))$ which is updated with every minibatch of $\mathbf{x} \sim p(\mathbf{x})$. The gradient for the current minibatch can be computed by treating the previous estimate of $\text{Cov}_{p(\mathbf{x})}(\boldsymbol{\mu}_\phi(\mathbf{x}))$ as constant.

2.3 Comparison with β-VAE

Recently proposed β-VAE (Higgins et al., 2017) proposes to modify the ELBO by upweighting the $\text{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))$ term in order to encourage the inference of disentangled factors:

$$\max_{\theta,\phi} \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \beta\,\text{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))\right], \tag{7}$$

where β is taken to be great than 1. Higher β is argued to encourage disentanglement at the cost of reconstruction error (the likelihood term in the ELBO). Authors report empirical results with β ranging from 4 to 250 depending on the dataset. As already mentioned, most VAE models proposed in the literature, including β-VAE, work with $N(\mathbf{0}, \mathbf{I})$ as the prior $p(\mathbf{z})$ and $N(\boldsymbol{\mu}_\phi(\mathbf{x}), \boldsymbol{\Sigma}_\phi(\mathbf{x}))$ with diagonal $\boldsymbol{\Sigma}_\phi(\mathbf{x})$ as the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$. This reduces the objective (7) to

$$\max_{\theta,\phi} \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \frac{\beta}{2}\left(\sum_i \left([\boldsymbol{\Sigma}_\phi(\mathbf{x})]_{ii} - \ln\left[\boldsymbol{\Sigma}_\phi(\mathbf{x})\right]_{ii}\right) + \|\boldsymbol{\mu}_\phi(\mathbf{x})\|_2^2\right)\right]. \tag{8}$$

For high values of β, β-VAE would try to pull $\boldsymbol{\mu}_\phi(\mathbf{x})$ towards zero and $\boldsymbol{\Sigma}_\phi(\mathbf{x})$ towards the identity matrix (as the minimum of $x - \ln x$ for $x > 0$ is at $x = 1$), thus making the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ insensitive to the observations. This is also reflected in the quality of the generated samples which is worse than VAE ($\beta = 1$), particularly for high values of β. Our proposed method does not have such increased tension between the likelihood term and the disentanglement objective, and the sample quality with our method is on par with the VAE.

Finally, we note that both β-VAE and our proposed method encourage disentanglement of inferred factors by pulling $\text{Cov}_{q_\phi(\mathbf{z})}(\mathbf{z})$ in Eq. (5) towards the identity matrix: β-VAE attempts to do it by making $\text{Cov}_{q_\phi(\mathbf{z}|\mathbf{x})}(\mathbf{z})$ close to $\mathbf{I}$ and $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}(\mathbf{z})$ close to $\mathbf{0}$ individually for all observations $\mathbf{x}$, while the proposed method directly works on $\text{Cov}_{q_\phi(\mathbf{z})}(\mathbf{z})$ (marginalizing over the observations $\mathbf{x}$) which retains the sensitivity of $q_\phi(\mathbf{z}|\mathbf{x})$ to the conditioned-upon observation.

3 Experiments

We evaluate our proposed method, referred as DIP-VAE subsequently (Disentangled Inferred Prior), on three datasets – (i) **CelebA** (Liu et al., 2015): It consists of $202,599$ RGB face images of celebrities. We use $64 \times 64 \times 3$ cropped images as used in several earlier works, using 90% for training and 10% for test. (ii) **3D Chairs** (Aubry et al., 2014): It consists of 1393 chair CAD models, with each model rendered from 31 azimuth angles and 2 elevation angles. Following earlier work (Yang et al., 2015; Dosovitskiy et al., 2015) that ignores near-duplicates, we use a subset of 809

Table 1: Disentanglement metric score Higgins et al. (2017) and reconstruction error (per pixel) on the test sets for 2D Shapes and CelebA ($\beta_1 = 4, \beta_2 = 20$ for 2D Shapes, and $\beta_1 = 4, \beta_2 = 8$ for CelebA)

Method	2D Shapes		CelebA	
	Metric	Reconst. error	Metric	Reconst. error
VAE	81.3	**0.0017**	7.5	**0.0876**
β-VAE (β=β_1)	80.7	0.0031	8.1	0.0937
β-VAE (β=β_2)	88.0	0.0076	7.1	0.1065
DIP-VAE	**98.1**	0.0018	**11.31**	0.0911

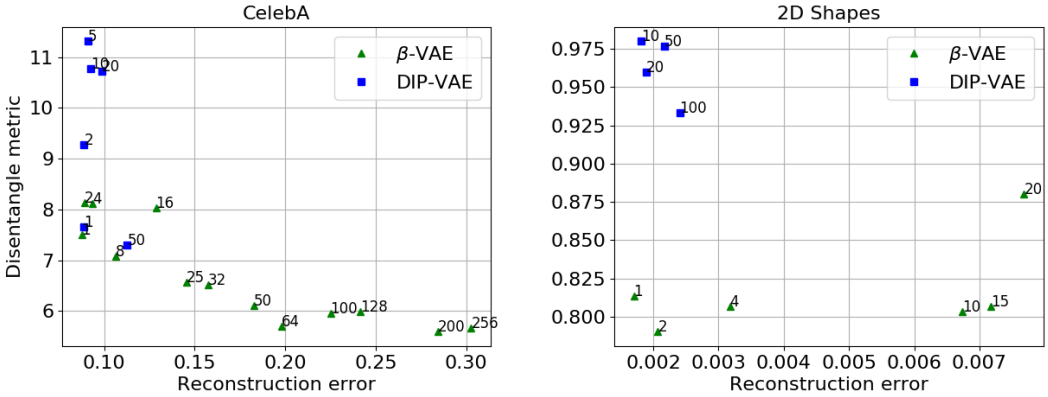


Figure 1: Disentanglement metric score (Higgins et al., 2017) as a function of reconstruction error for β-VAE and the proposed DIP-VAE (*left:* CelebA, *right:* 2D Shapes). The plots are generated by varying β for β-VAE and λ_{od} for DIP-VAE with λ_d set to $10\lambda_{od}$.

chair models in our experiments. We use the binary masks of the chairs as the observed data in our experiments following (Higgins et al., 2017). First 80% of the models are used for training and the rest are used for test. **(iii) 2D Shapes** (Higgins et al., 2017): This is a synthetic dataset of binary 2D shapes generated from the Cartesian product of the shape (heart, oval and square), x-position (32 values), y-position (32 values), scale (6 values) and rotation (40 values). We consider two baselines for the task of unsupervised inference of disentangled factors: (i) VAE (Kingma & Welling, 2013; Rezende et al., 2014), and (ii) the recently proposed β-VAE (Higgins et al., 2017). To be consistent with the evaluations in (Higgins et al., 2017), we use the same CNN network architectures (for our encoder and decoder), and same latent dimensions as used in (Higgins et al., 2017) for CelebA, 3D Chairs, 2D Shapes datasets.

Hyperparameters. For the proposed DIP-VAE, in all our experiments we vary λ_d in the set $\{1, 2, 5, 10, 20, 50\}$ while fixing $\lambda_d = 10\lambda_{od}$. For β-VAE, we experiment with $\beta = \{1, 2, 4, 8, 16, 25, 32, 64, 100, 128, 200, 256\}$ (where $\beta = 1$ corresponds to VAE). For both CelebA and 2D Shapes, we show the results for the best performing models in terms of the *disentanglement metric score* which was introduced in (Higgins et al., 2017). For 3D Chairs data, this metric is at 100% for almost all models and we pick the models based on our subjective evaluation of the reconstruction quality and disentanglement.

Disentanglement metric score and reconstruction error. Higgins et al. (2017) propose a metric to evaluate the disentanglement performance of the inference mechanism. It assumes that the ground truth generative factors are available. It works by first sampling a generative factor y, followed by sampling L pairs of examples such that for each pair, the sampled generative factor takes the same value. Given the inferred $\mathbf{z}_x := \boldsymbol{\mu}_\phi \mathbf{x}$ for each example $\mathbf{x}$, they compute the absolute difference of these vectors for each pair, followed by averaging these difference vectors. This average difference vector is assigned the label of y. By sampling n such minibatches of L pairs, we get n such averaged difference vectors for the factor y. This process is repeated for all generative factors. A low capacity multiclass classifier is then trained on these vectors to predict the identity of the corresponding

Table 2: Attribute classification accuracy on CelebA: A classifier $\mathbf{w}^k = \frac{1}{|\mathbf{x}_i : y_i^k = 1|} \sum_{\mathbf{x}_i : y_i^k = 1} \mu_\phi(\mathbf{x}_i) - \frac{1}{|\mathbf{x}_i : y_i^k = 0|} \sum_{\mathbf{x}_i : y_i^k = 0} \mu_\phi(\mathbf{x}_i)$ is computed for every attribute k using the training set and a bias is learned by minimizing the hinge loss. Accuracy on other attributes stays about same across all methods.

Method	Arched Eyebrows	Attractive	Bangs	Black hair	Blond hair	Heavy makeup	Male	Mouth slighly open	No Beard	Wavy hair	Wearing hat	Wearing lipstick
VAE	71.8	73.0	89.8	78.0	88.9	79.6	83.9	**76.3**	**87.3**	70.2	95.8	83.0
$\beta = 2$	71.6	72.6	90.6	79.3	89.1	79.3	83.5	76.1	86.9	67.8	95.9	82.4
$\beta = 4$	71.6	72.6	90.0	76.6	88.9	77.8	82.3	75.7	85.3	66.8	95.8	80.6
$\beta = 8$	71.6	71.7	90.0	76.0	87.2	76.2	80.5	73.1	85.3	63.7	95.8	79.6
DIP-VAE	**73.7**	**73.2**	**90.9**	**80.6**	**91.9**	**81.5**	**85.9**	75.9	85.3	**71.5**	**96.2**	**84.7**

generative factor. In all our experiments, we use a one-vs-rest linear SVM with weight on the hinge loss C set to 0.01 and weight on the regularizer set to 1. Higgins et al. (2017) argue that this metric captures the disentangled property of the inferred latents reasonably well. Table 1 shows the disentanglement metric scores along with reconstruction error (which directly corresponds to the data likelihood) for the test sets for CelebA and 2D Shapes data. It is evident that the proposed DIP-VAE outperforms β-VAE both in terms of the disentanglement metric score and reconstruction error. Further we also show the plot of how the disentanglement metric changes with the reconstruction error as we vary the hyperparameter for both methods (β and λ_{od}, respectively). It is clear that the proposed method gives much higher disentanglement metric score at little to no cost on the reconstruction error when compared VAE ($\beta = 1$). The reconstruction error for β-VAE gets much worse as β is increased.

Binary attribute Classification for CelebA. We also experiment with predicting the binary attribute values for each test example in CelebA from the inferred $\boldsymbol{\mu}_\phi(\mathbf{x})$. For each attribute k, we compute the *attribute vector* $\mathbf{w}^k = \frac{1}{|\mathbf{x}_i : y_i^k = 1|} \sum_{\mathbf{x}_i : y_i^k = 1} \mu_\phi(\mathbf{x}_i) - \frac{1}{|\mathbf{x}_i : y_i^k = 0|} \sum_{\mathbf{x}_i : y_i^k = 0} \mu_\phi(\mathbf{x}_i)$ from the training set, and project the $\boldsymbol{\mu}_\phi(\mathbf{x})$ along these vectors. A bias is learned on these scalars (by minimizing hinge loss) which is then used for classifying the test examples. Table 2 shows the results for the attribute which show the highest change across various methods (most other attribute accuracies do not change). The proposed DIP-VAE outperforms both VAE and β-VAE for most attributes. The performance of β-VAE gets worse as β is increased further.

Correlations in the inferred latent space. We visualize the Pearson's correlations between dimensions of the inferred latents $\boldsymbol{\mu}_\phi(\mathbf{x})$. We also visualize the correlations between inferred latent dimensions and the ground truth attributes which can be taken as a proxy for true generative factors. Tables 3 and 4 show these correlations for VAE, β-VAE and the proposed DIP-VAE. We observe less correlations between inferred latents for DIP-VAE. Latent correlations for β-VAE are even higher than those for VAE which seems to be going against the objective of disentanglement. This can be due to the fact that β-VAE give less weight to the ELBO which implies the $\mathrm{KL}(q_\phi(\mathbf{z}) \| p(\mathbf{z}))$ does not get minimized as well (see Eq. (3)). Indeed Eq. (8) indicates that β-VAE gives more importance to minimizing the ℓ_2-norm of individual posterior means $\boldsymbol{\mu}_\phi(\mathbf{x})$ and does not minimize their correlations as DIP-VAE.

To further analyze the latent correlations with the ground truth attributes, we pick one attribute at a time, and pick the latent dimension that has highest correlation with that attribute. Then we plot the correlations of this latent dimension with rest of the attributes (picking top 4 correlations). This will indicate the entanglement of a latent dimension with the ground truth attributes. It should be noted that many attributes in CelebA are naturally entangled (e.g., lipstick, heavy-makeup and gender) which can be quite difficult to disentangle. These plots for two of the attributes are shown in Fig. 2.

Table 3: Correlations between inferred latents *(top)*, and correlations between ground truth attributes and inferred latents *(bottom)* for 2D Shapes data (Higgins et al., 2017).

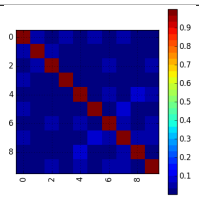
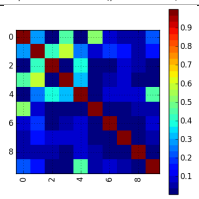
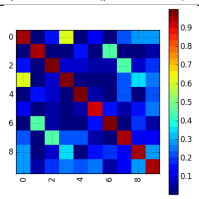
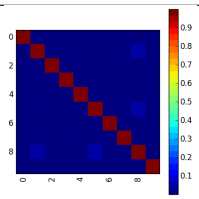
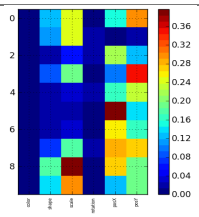
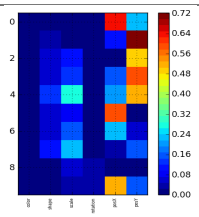
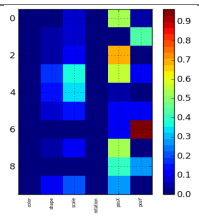
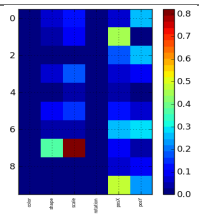
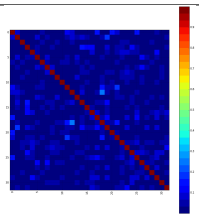
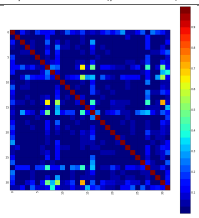
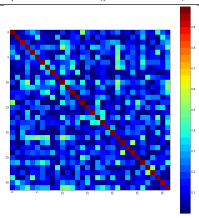
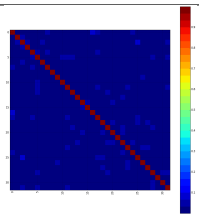
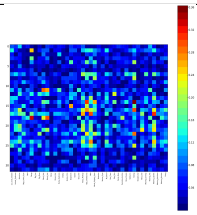
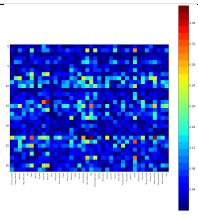
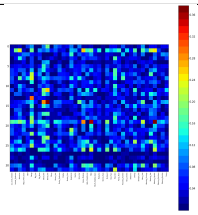
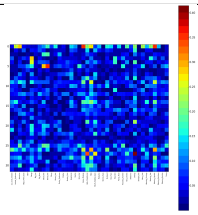
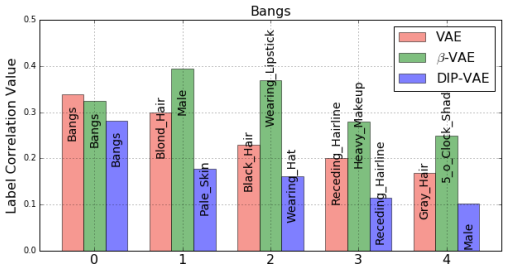
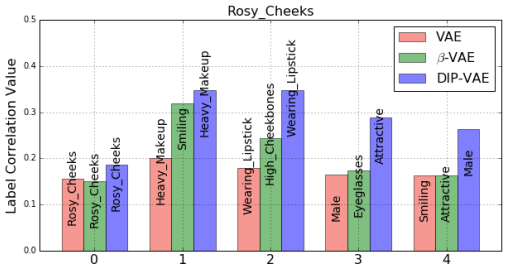
	VAE	β-VAE ($\beta = 4$)	β-VAE ($\beta = 20$)	DIP-VAE
Latents auto-correlations				
Latents-attributes correlations				

Table 4: Correlations between inferred latents *(top)*, and correlations between ground truth attributes and inferred latents *(bottom)* for CelebA dataset.

	VAE	β-VAE ($\beta = 8$)	β-VAE ($\beta = 64$)	DIP-VAE
Latents auto-correlations				
Latents-attributes correlations				



(a) *Bangs*



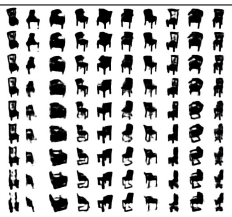
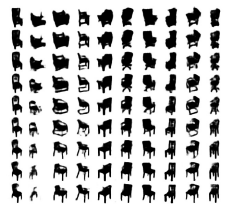
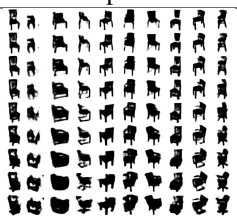
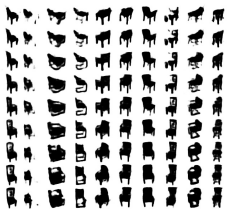
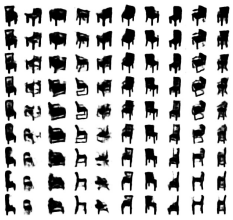
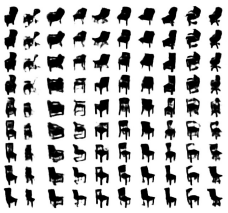
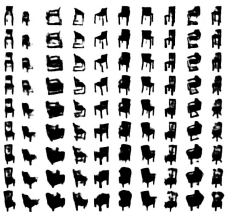
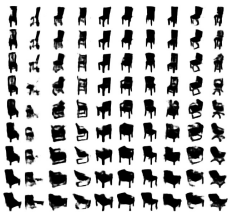
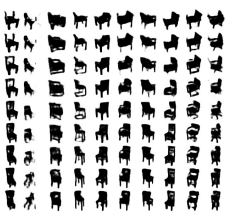
(b) *Rosy Cheeks*

Figure 2: Top attributes similar to *Bangs* and *Rosy Cheeks* attribute in CelebA dataset based on correlations with inferred latents for VAE, β-VAE and the proposed DIP-VAE, respectively.

Table 5: Qualitative results for disentanglement in CelebA dataset.

	VAE	β-VAE ($\beta = 4$)	Proposed
Smile			
Azimuth			
Hair color			

Table 6: Qualitative results for disentanglement in Chairs dataset.

	VAE	β-VAE ($\beta = 4$)	Proposed
Leg Shape			
Rotation			
Size			

4 DISCUSSION

We proposed a principled variational framework with infer disentangled latents from unlabeled observations. Unlike β-VAE our variational objective does not have any conflict between the data log-likelihood and the disentanglement of the inferred latents, which is reflected in our empirical results that outperform β-VAE.

REFERENCES

Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. In *NIPS 2016 Workshop on Adversarial Training. In review for ICLR*, volume 2016, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

Mathieu Aubry, Daniel Maturana, Alexei A Efros, Bryan C Russell, and Josef Sivic. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3762–3769, 2014.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 2017.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2172–2180, 2016.

Brian Cheung, Jesse A Livezey, Arjun K Bansal, and Bruno A Olshausen. Discovering hidden factors of variation in deep networks. *arXiv preprint arXiv:1412.6583*, 2014.

Taco Cohen and Max Welling. Learning the irreducible representations of commutative lie groups. In *International Conference on Machine Learning*, pp. 1755–1763, 2014.

Taco S Cohen and Max Welling. Transformation properties of learned visual representations. In *International Conference on Learning Representations*, 2015.

Emily Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. *arXiv preprint arXiv:1705.10915*, 2017.

Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

Alexey Dosovitskiy, Jost Tobias Springenberg, and Thomas Brox. Learning to generate chairs with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1538–1546, 2015.

Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martin Arjovsky, Olivier Mastropietro, and Aaron Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.

Gintare Karolina Dziugaite, Daniel M Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pp. 258–267. AUAI Press, 2015.

Marco Fraccaro, Simon Kamronn, Ulrich Paquet, and Ole Winther. A disentangled recognition and nonlinear dynamics model for unsupervised learning. *arXiv preprint arXiv:1710.05741*, 2017.

Michael C Fu. Gradient estimation. *Handbooks in operations research and management science*, 13: 575–616, 2006.

Samuel Gershman and Noah Goodman. Amortized inference in probabilistic reasoning. In *Proceedings of the Cognitive Science Society*, volume 36, 2014.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Ross Goroshin, Michael F Mathieu, and Yann LeCun. Learning to linearize under uncertainty. In *Advances in Neural Information Processing Systems*, pp. 1234–1242, 2015.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

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.

Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. Transforming auto-encoders. In *International Conference on Artificial Neural Networks*, pp. 44–51. Springer, 2011.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Xianxu Hou, Linlin Shen, Ke Sun, and Guoping Qiu. Deep feature consistent variational autoencoder. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pp. 1133–1141. IEEE, 2017.

Wei-Ning Hsu, Yu Zhang, and James Glass. Unsupervised learning of disentangled latent representations from sequential data. In *Advances in neural information processing systems*, 2017.

Theofanis Karaletsos, Serge Belongie, and Gunnar Rätsch. Bayesian representation learning with oracle constraints. *arXiv preprint arXiv:1506.05011*, 2015.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *Advances in Neural Information Processing Systems*, pp. 2539–2547, 2015.

Abhishek Kumar, Prasanna Sattigeri, and P Thomas Fletcher. Improved semi-supervised learning with gans using manifold invariances. *arXiv preprint arXiv:1705.08850*, 2017.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, pp. 1–101, 2016.

Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pp. 1718–1727, 2015.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. Disentangling factors of variation in deep representation using adversarial training. In *Advances in Neural Information Processing Systems*, pp. 5040–5048, 2016.

Yishu Miao, Lei Yu, and Phil Blunsom. Neural variational inference for text processing. In *International Conference on Machine Learning*, pp. 1727–1736, 2016.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, pp. 271–279, 2016.

Scott Reed, Kihyuk Sohn, Yuting Zhang, and Honglak Lee. Learning to disentangle factors of variation with manifold interaction. In *International Conference on Machine Learning*, pp. 1431–1439, 2014.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Karl Ridgeway. A survey of inductive biases for factorial representation-learning. *arXiv preprint arXiv:1612.05299*, 2016.

N Siddharth, Brooks Paige, Van de Meent, Alban Desmaison, Frank Wood, Noah D Goodman, Pushmeet Kohli, Philip HS Torr, et al. Learning disentangled representations with semi-supervised deep generative models. *arXiv preprint arXiv:1706.00400*, 2017.

Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On integral probability metrics,\phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

Andreas Stuhlmüller, Jacob Taylor, and Noah Goodman. Learning stochastic inverses. In *Advances in neural information processing systems*, pp. 3048–3056, 2013.

Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

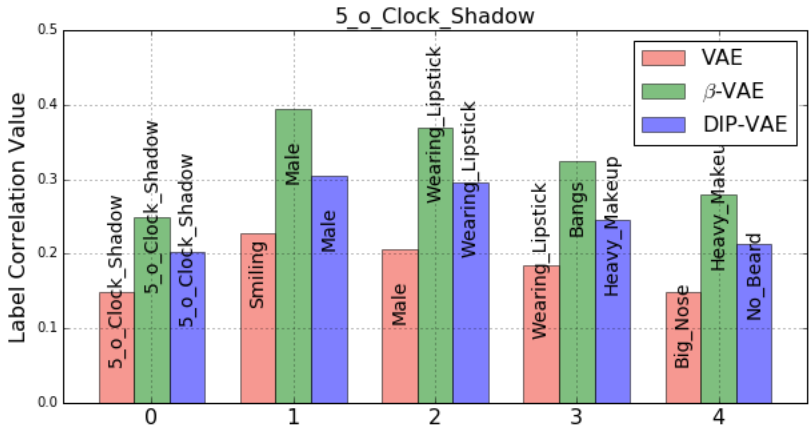
William F Whitney, Michael Chang, Tejas Kulkarni, and Joshua B Tenenbaum. Understanding visual concepts with continuation learning. *arXiv preprint arXiv:1602.06822*, 2016.

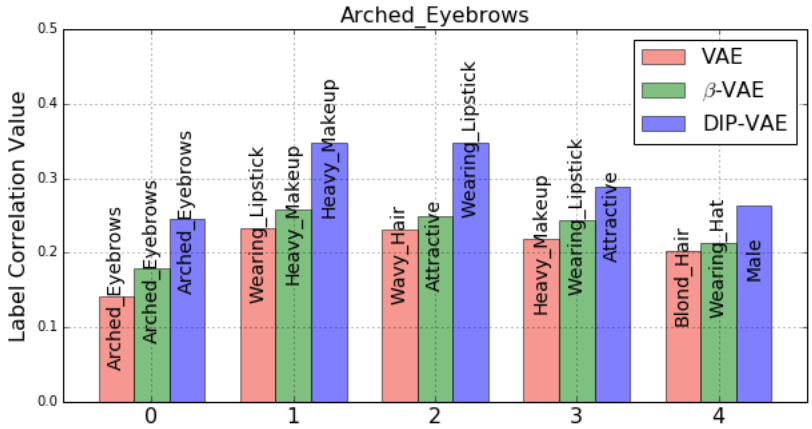
Jimei Yang, Scott E Reed, Ming-Hsuan Yang, and Honglak Lee. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. In *Advances in Neural Information Processing Systems*, pp. 1099–1107, 2015.

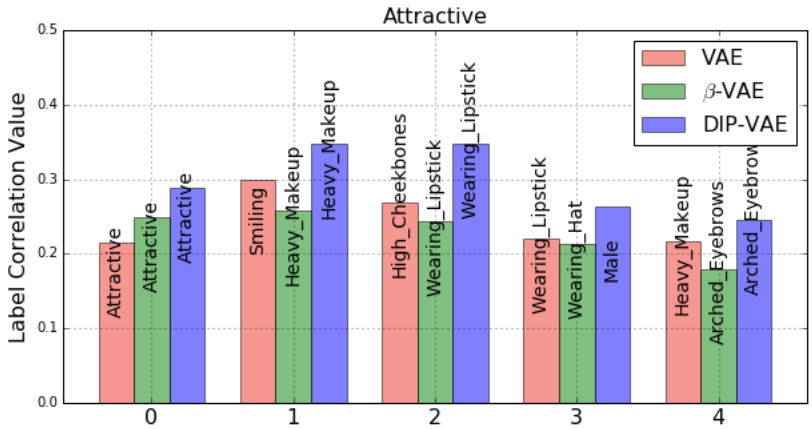
Shengjia Zhao, Jiaming Song, and Stefano Ermon. Towards deeper understanding of variational autoencoding models. *arXiv preprint arXiv:1702.08658*, 2017.

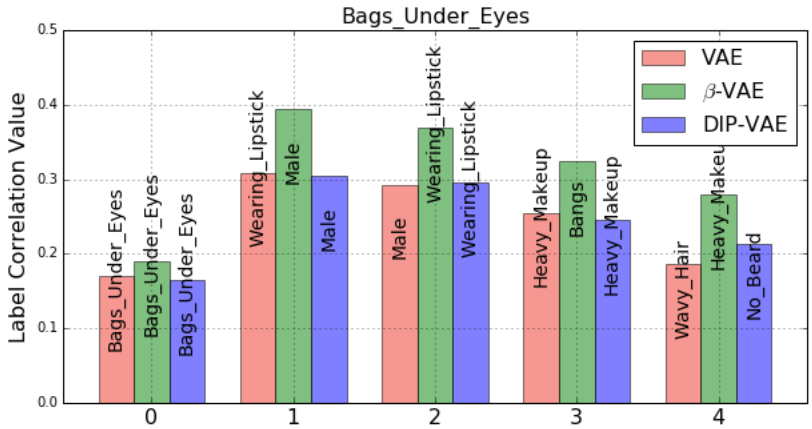
Appendix

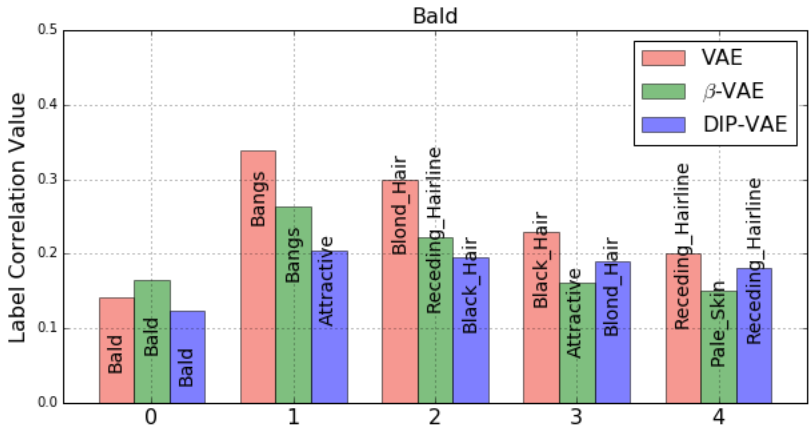
A CORRELATIONS BETWEEN INFERRED LATENT FACTORS AND GROUND TRUTH ATTRIBUTES FOR CELEBA

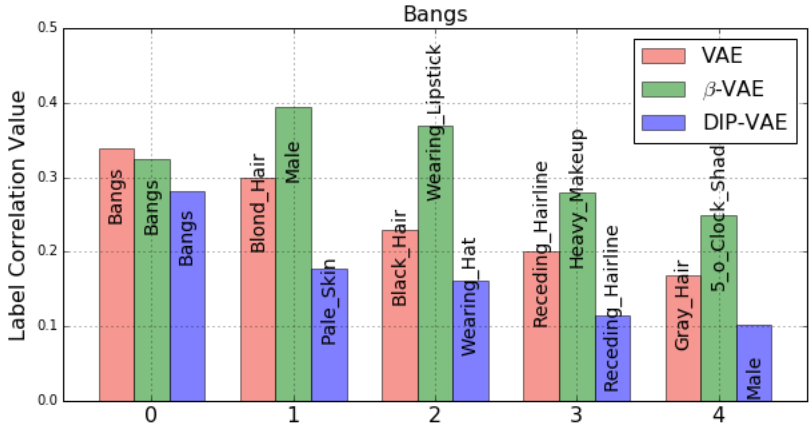


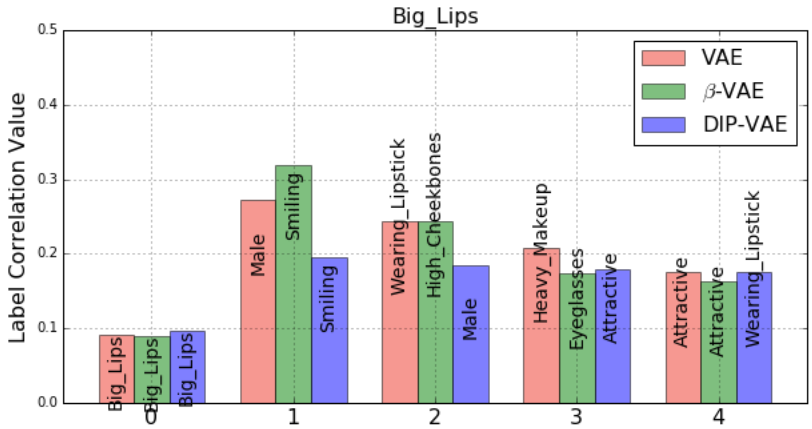


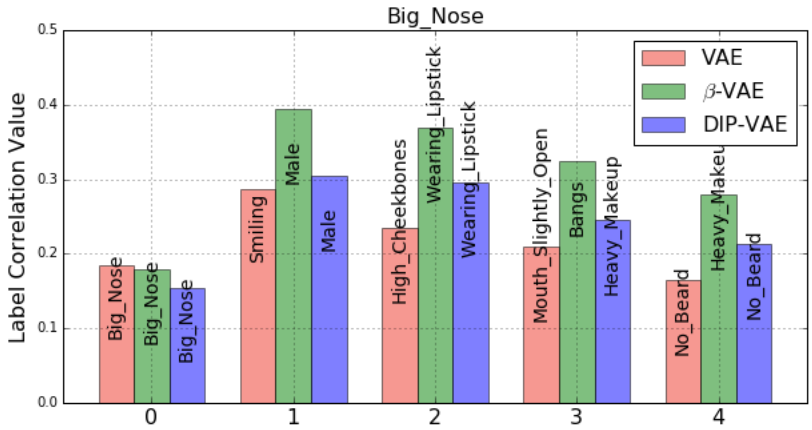


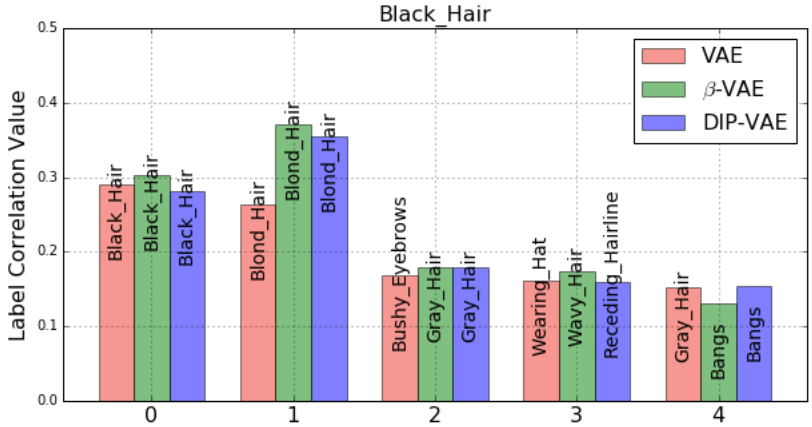

Bags_Under_Eyes
Label Correlation Value
0.0
0.1
0.2
0.3
0.4
0.5
VAE
β-VAE
DIP-VAE
Bags_Under_Eyes
Bags_Under_Eyes
Bags_Under_Eyes
Wearing_Lipstick
Male
Male
Male
Wearing_Lipstick
Wearing_Lipstick
Heavy_Makeup
Bangs
Heavy_Makeup
Wavy_Hair
Heavy_Makeu
No_Beard
0
1
2
3
4

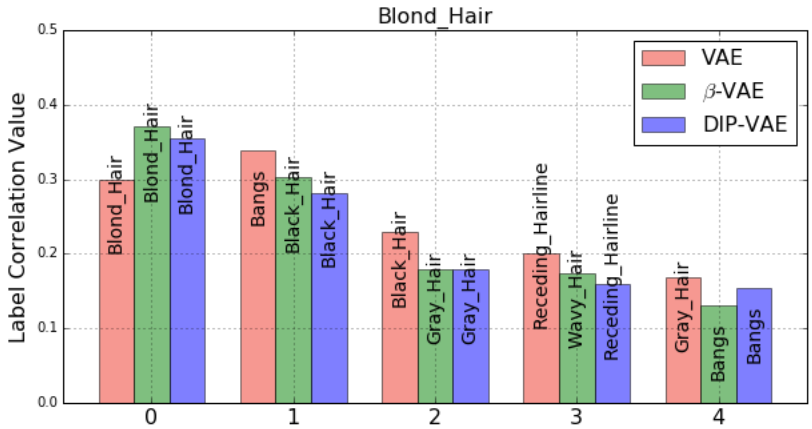

Bald
Label Correlation Value
0.0
0.1
0.2
0.3
0.4
0.5
VAE
β-VAE
DIP-VAE
Bald
Bald
Bald
Bangs
Bangs
Attractive
Blond_Hair
Receding_Hairline
Black_Hair
Black_Hair
Attractive
Blond_Hair
Receding_Hairline
Pale_Skin
Receding_Hairline
0
1
2
3
4

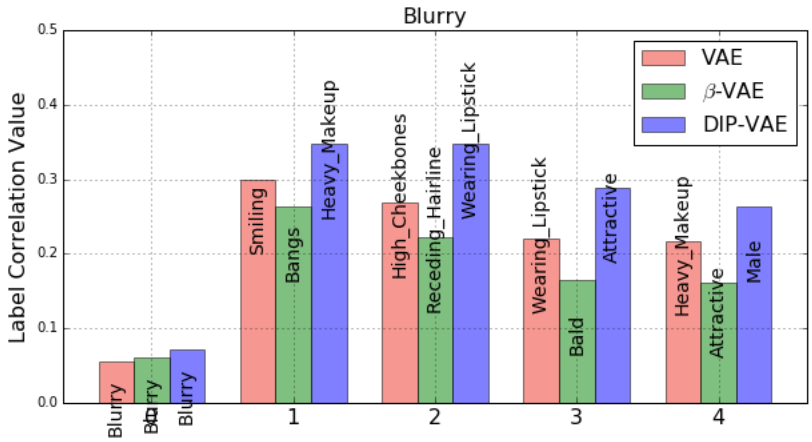

Bangs
Label Correlation Value
0.0
0.1
0.2
0.3
0.4
0.5
VAE
β-VAE
DIP-VAE
Bangs
Bangs
Bangs
Blond_Hair
Male
Pale_Skin
Black_Hair
Wearing_Lipstick
Wearing_Hat
Receding_Hairline
Heavy_Makeup
Receding_Hairline
Gray_Hair
5_o_Clock_Shad
Male
0
1
2
3
4

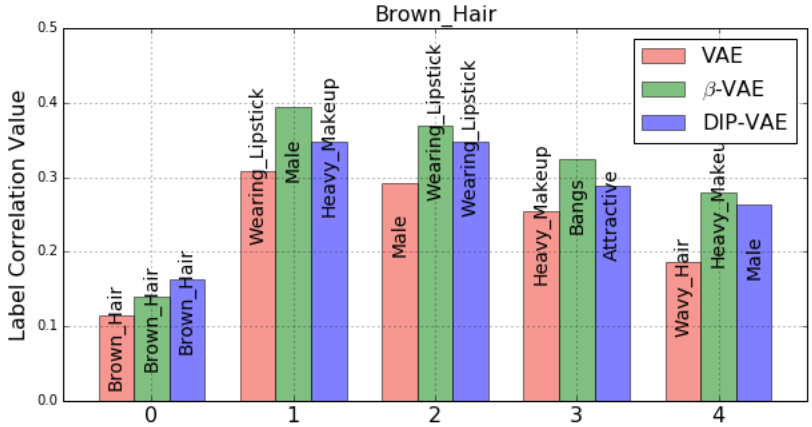

Big_Lips
Label Correlation Value
0.0
0.1
0.2
0.3
0.4
0.5
VAE
β-VAE
DIP-VAE
Big_Lips
Big_Lips
Big_Lips
Male
Smiling
Smiling
Wearing_Lipstick
High_Cheekbones
Male
Heavy_Makeup
Eyeglasses
Attractive
Attractive
Attractive
Wearing_Lipstick
0
1
2
3
4

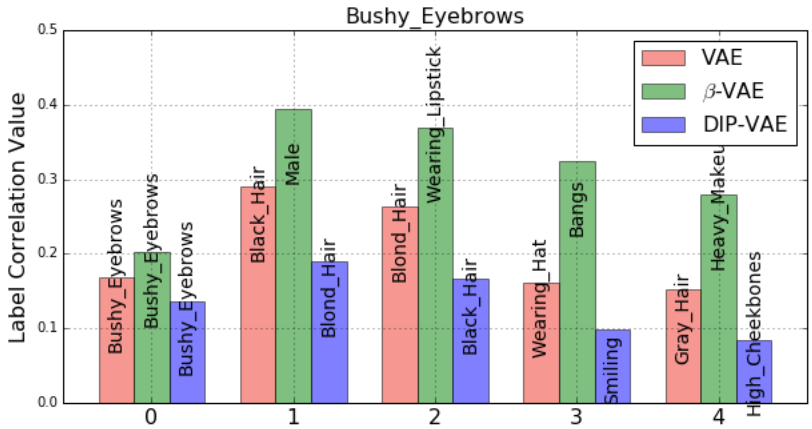

Big_Nose
Label Correlation Value
0.0
0.1
0.2
0.3
0.4
0.5
VAE
β-VAE
DIP-VAE
Big_Nose
Big_Nose
Big_Nose
Smiling
Male
Male
High_Cheekbones
Wearing_Lipstick
Wearing_Lipstick
Mouth_Slightly_Open
Bangs
Heavy_Makeup
No_Beard
Heavy_Makeup
No_Beard
0
1
2
3
4

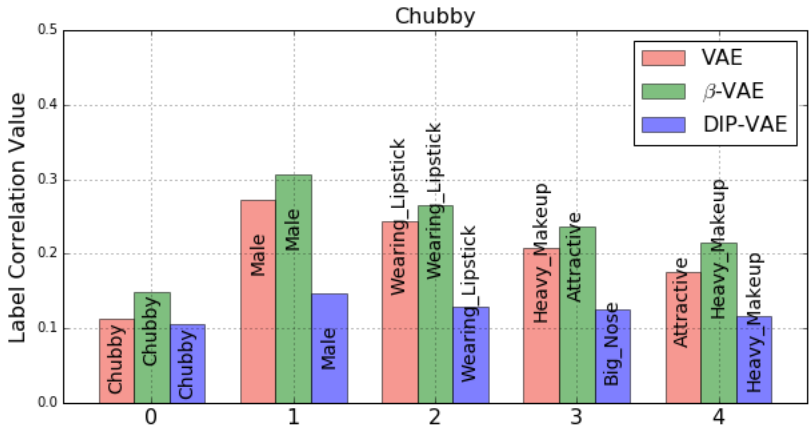

Black_Hair
Label Correlation Value
0.0
0.1
0.2
0.3
0.4
0.5
VAE
β-VAE
DIP-VAE
Black_Hair
Black_Hair
Black_Hair
Blond_Hair
Blond_Hair
Blond_Hair
Bushy_Eyebrows
Gray_Hair
Gray_Hair
Wearing_Hat
Wavy_Hair
Receding_Hairline
Gray_Hair
Bangs
Bangs
0
1
2
3
4

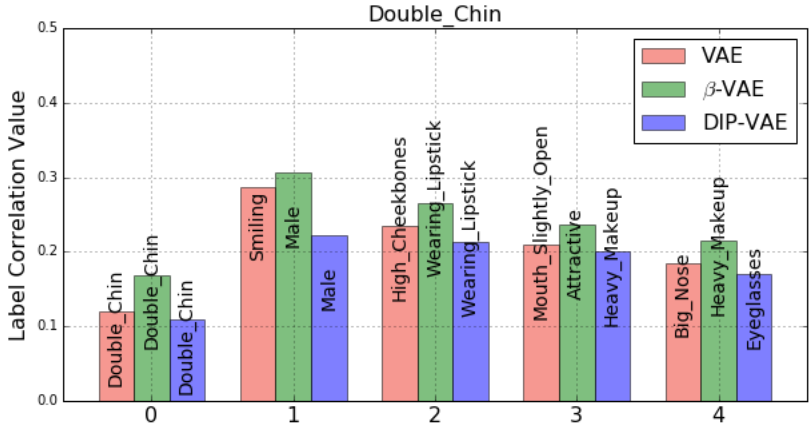

Blond_Hair
Label Correlation Value
VAE
β-VAE
DIP-VAE
0.0
0.1
0.2
0.3
0.4
0.5
0
1
2
3
4
Blond_Hair
Blond_Hair
Blond_Hair
Bangs
Black_Hair
Black_Hair
Black_Hair
Gray_Hair
Gray_Hair
Receding_Hairline
Wavy_Hair
Receding_Hairline
Gray_Hair
Bangs
Bangs

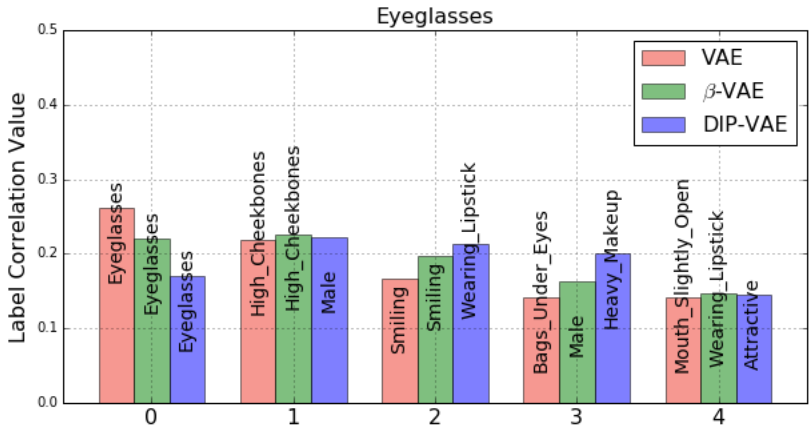

Blurry
Label Correlation Value
VAE
β-VAE
DIP-VAE
0.0
0.1
0.2
0.3
0.4
0.5
1
2
3
4
Blurry
Blurry
Smiling
Bangs
Heavy_Makeup
High_Cheekbones
Receding_Hairline
Wearing_Lipstick
Wearing_Lipstick
Bald
Attractive
Heavy_Makeup
Attractive
Male

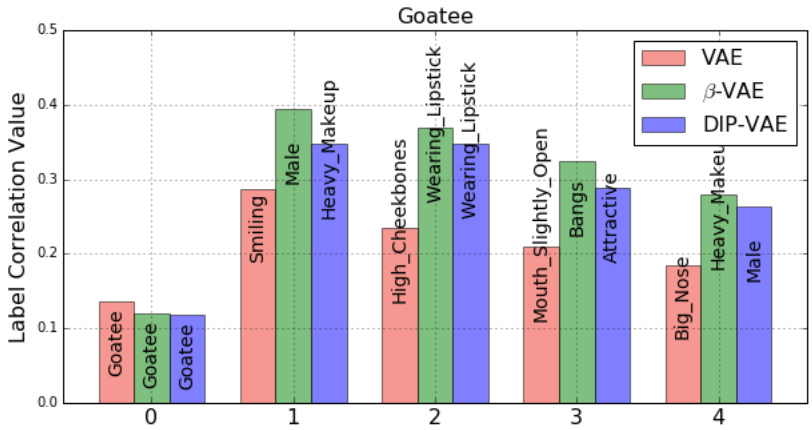

Brown_Hair
Label Correlation Value
VAE
β-VAE
DIP-VAE
0.0
0.1
0.2
0.3
0.4
0.5
0
1
2
3
4
Brown_Hair
Brown_Hair
Brown_Hair
Wearing_Lipstick
Male
Heavy_Makeup
Male
Wearing_Lipstick
Wearing_Lipstick
Heavy_Makeup
Bangs
Attractive
Wavy_Hair
Male

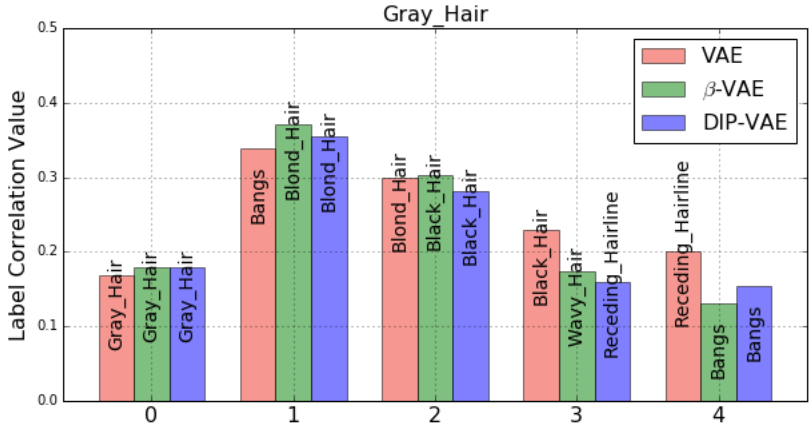

Bushy_Eyebrows
Label Correlation Value
0.0
0.1
0.2
0.3
0.4
0.5
VAE
β-VAE
DIP-VAE
Bushy_Eyebrows
Bushy_Eyebrows
Bushy_Eyebrows
Black_Hair
Male
Blond_Hair
Blond_Hair
Wearing_Lipstick
Black_Hair
Wearing_Hat
Bangs
Smiling
Gray_Hair
Heavy_Makeup
High_Cheekbones
0
1
2
3
4

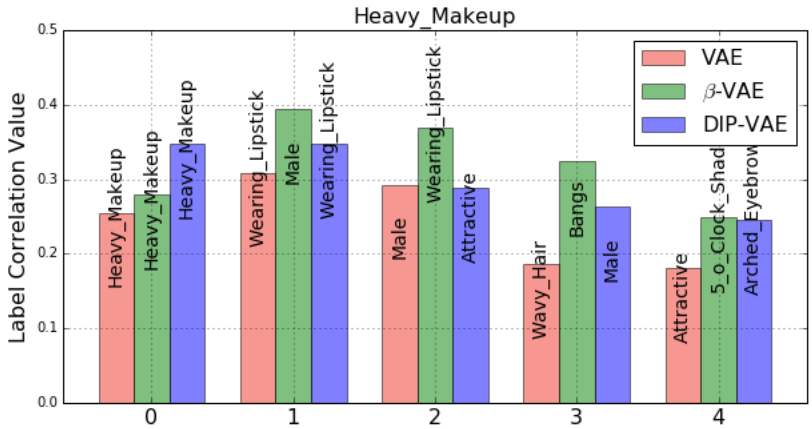

Chubby
Label Correlation Value
0.0
0.1
0.2
0.3
0.4
0.5
VAE
β-VAE
DIP-VAE
Chubby
Chubby
Chubby
Male
Male
Male
Wearing_Lipstick
Wearing_Lipstick
Wearing_Lipstick
Heavy_Makeup
Attractive
Big_Nose
Attractive
Heavy_Makeup
Heavy_Makeup
0
1
2
3
4

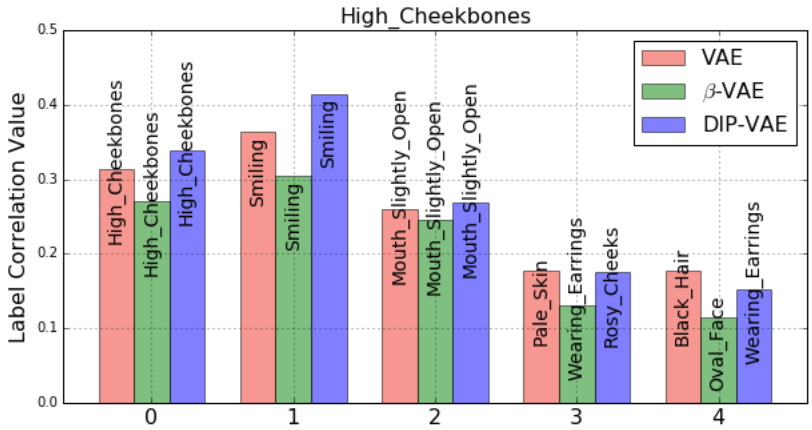

Double_Chin
Label Correlation Value
0.0
0.1
0.2
0.3
0.4
0.5
VAE
β-VAE
DIP-VAE
Double_Chin
Double_Chin
Double_Chin
Smiling
Male
Male
High_Cheekbones
Wearing_Lipstick
Wearing_Lipstick
Mouth_Slightly_Open
Attractive
Heavy_Makeup
Big_Nose
Heavy_Makeup
Eyeglasses
0
1
2
3
4

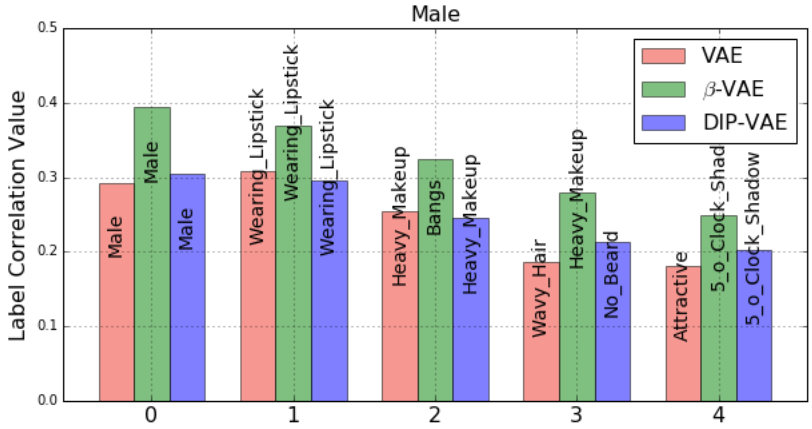

Eyeglasses
Label Correlation Value
0.0
0.1
0.2
0.3
0.4
0.5
VAE
β-VAE
DIP-VAE
Eyeglasses
Eyeglasses
Eyeglasses
High_Cheekbones
High_Cheekbones
Male
Smiling
Smiling
Wearing_Lipstick
Bags_Under_Eyes
Male
Heavy_Makeup
Mouth_Slightly_Open
Wearing_Lipstick
Attractive
0
1
2
3
4

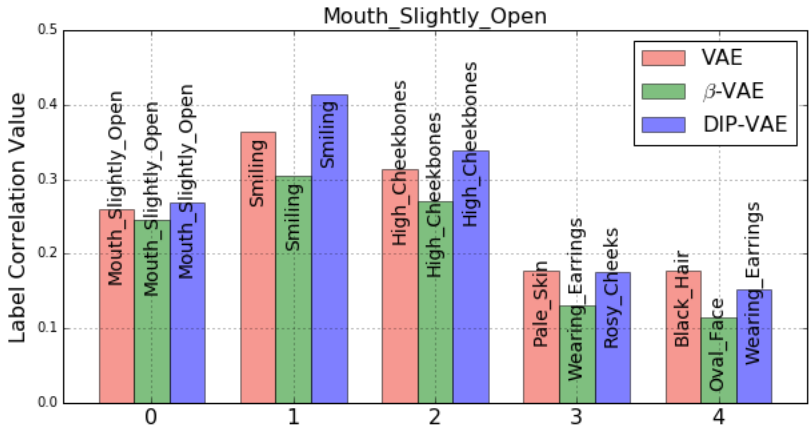

Goatee
Label Correlation Value
0.0
0.1
0.2
0.3
0.4
0.5
VAE
β-VAE
DIP-VAE
Goatee
Goatee
Goatee
Smiling
Male
Heavy_Makeup
High_Cheekbones
Wearing_Lipstick
Wearing_Lipstick
Mouth_Slightly_Open
Bangs
Attractive
Big_Nose
Male
0
1
2
3
4

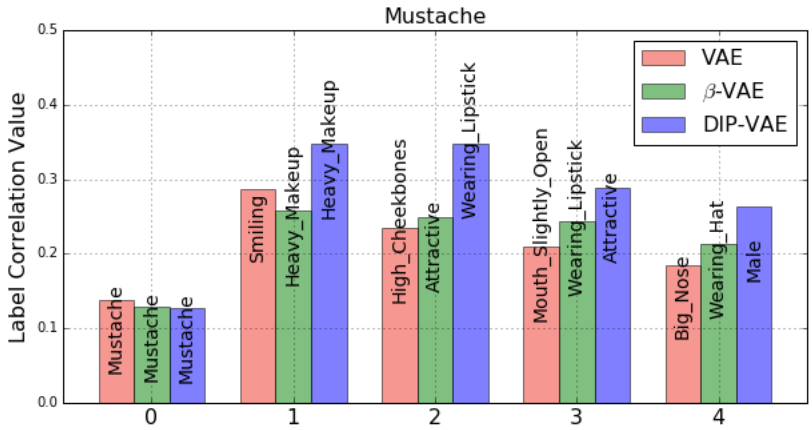

Gray_Hair
Label Correlation Value
0.0
0.1
0.2
0.3
0.4
0.5
VAE
β-VAE
DIP-VAE
Gray_Hair
Gray_Hair
Gray_Hair
Bangs
Blond_Hair
Blond_Hair
Blond_Hair
Black_Hair
Black_Hair
Black_Hair
Wavy_Hair
Receding_Hairline
Receding_Hairline
Bangs
Bangs
0
1
2
3
4

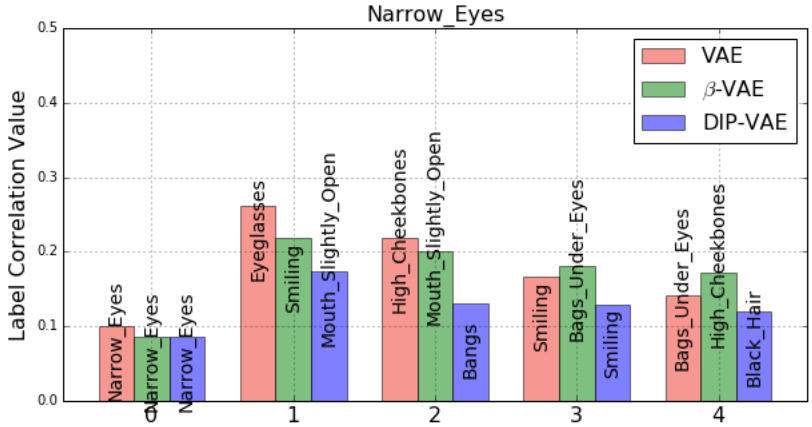

Heavy_Makeup
Label Correlation Value
VAE
β-VAE
DIP-VAE
Heavy_Makeup
Heavy_Makeup
Heavy_Makeup
Wearing_Lipstick
Male
Wearing_Lipstick
Male
Wearing_Lipstick
Attractive
Wavy_Hair
Bangs
Male
Attractive
5_o_Clock_Shad
Arched_Eyebrow
0
1
2
3
4

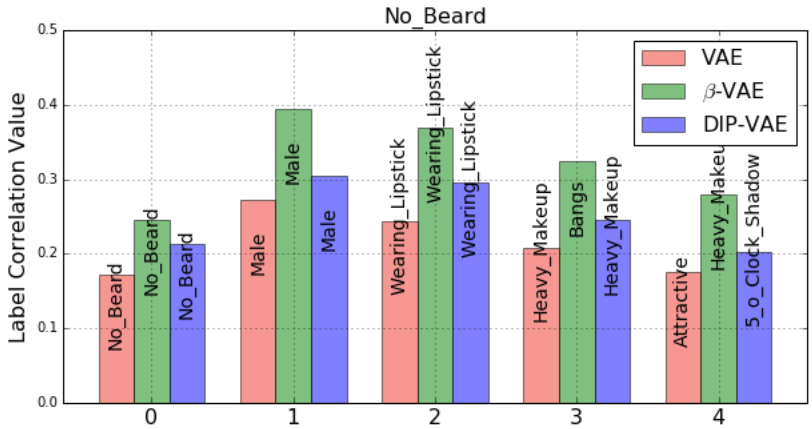

High_Cheekbones
Label Correlation Value
VAE
β-VAE
DIP-VAE
High_Cheekbones
High_Cheekbones
High_Cheekbones
Smiling
Smiling
Smiling
Mouth_Slightly_Open
Mouth_Slightly_Open
Mouth_Slightly_Open
Pale_Skin
Wearing_Earrings
Rosy_Cheeks
Black_Hair
Oval_Face
Wearing_Earrings
0
1
2
3
4

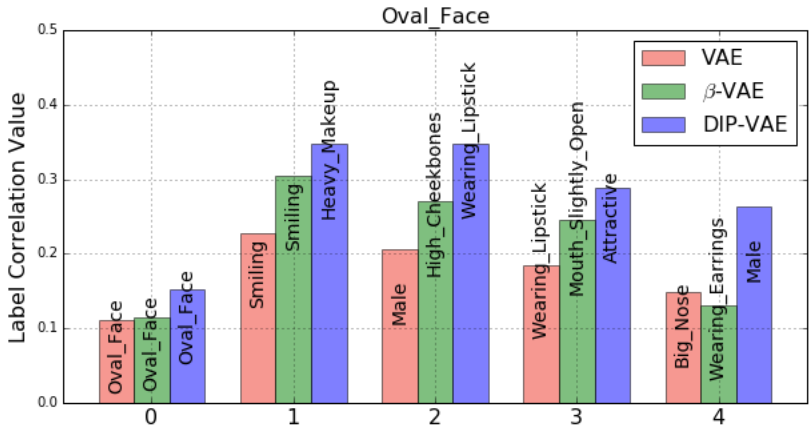

Male
Label Correlation Value
VAE
β-VAE
DIP-VAE
Male
Male
Male
Wearing_Lipstick
Wearing_Lipstick
Wearing_Lipstick
Heavy_Makeup
Bangs
Heavy_Makeup
Wavy_Hair
Heavy_Makeup
No_Beard
Attractive
5_o_Clock_Shad
5_o_Clock_Shadow
0
1
2
3
4

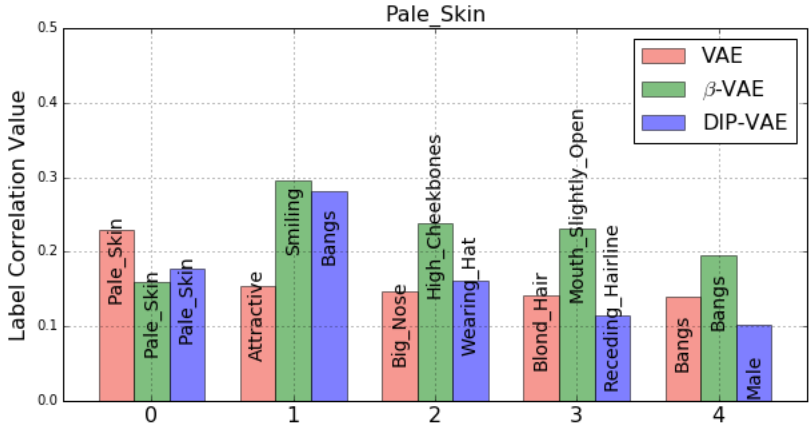

Mouth_Slightly_Open
Label Correlation Value
VAE
β-VAE
DIP-VAE
Mouth_Slightly_Open
Mouth_Slightly_Open
Mouth_Slightly_Open
Smiling
Smiling
Smiling
High_Cheekbones
High_Cheekbones
High_Cheekbones
Pale_Skin
Wearing_Earrings
Rosy_Cheeks
Black_Hair
Oval_Face
Wearing_Earrings
0
1
2
3
4

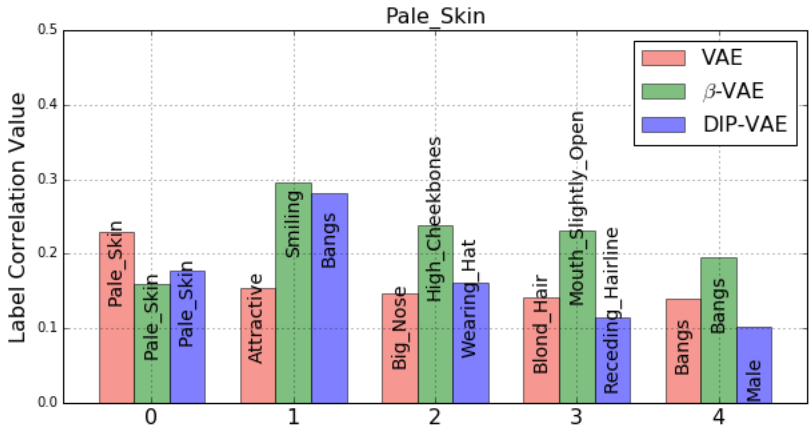

Mustache
Label Correlation Value
VAE
β-VAE
DIP-VAE
Mustache
Mustache
Mustache
Smiling
Heavy_Makeup
Heavy_Makeup
High_Cheekbones
Attractive
Wearing_Lipstick
Mouth_Slightly_Open
Wearing_Lipstick
Attractive
Big_Nose
Wearing_Hat
Male
0
1
2
3
4

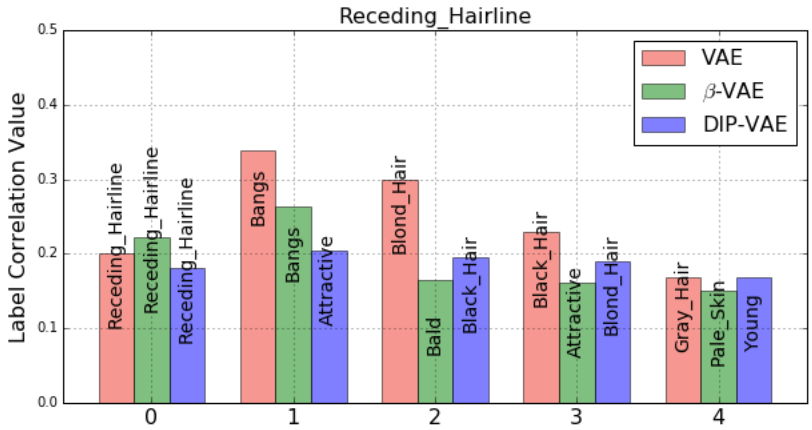

Narrow_Eyes
Label Correlation Value
VAE
β-VAE
DIP-VAE
Narrow_Eyes
Narrow_Eyes
Narrow_Eyes
Eyeglasses
Smiling
Mouth_Slightly_Open
High_Cheekbones
Mouth_Slightly_Open
Bangs
Smiling
Bags_Under_Eyes
Smiling
Bags_Under_Eyes
High_Cheekbones
Black_Hair
0
1
2
3
4

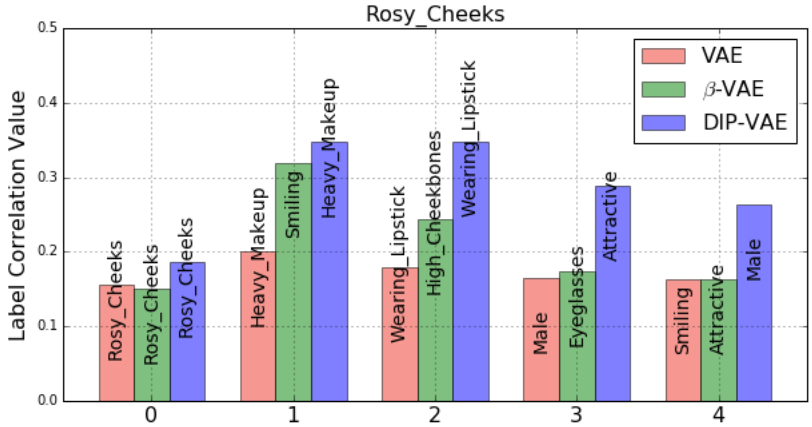

No_Beard
Label Correlation Value
VAE
β-VAE
DIP-VAE
No_Beard
No_Beard
No_Beard
Male
Male
Male
Wearing_Lipstick
Wearing_Lipstick
Wearing_Lipstick
Heavy_Makeup
Bangs
Heavy_Makeup
Attractive
Heavy_Makeup
5_o_Clock_Shadow
0
1
2
3
4

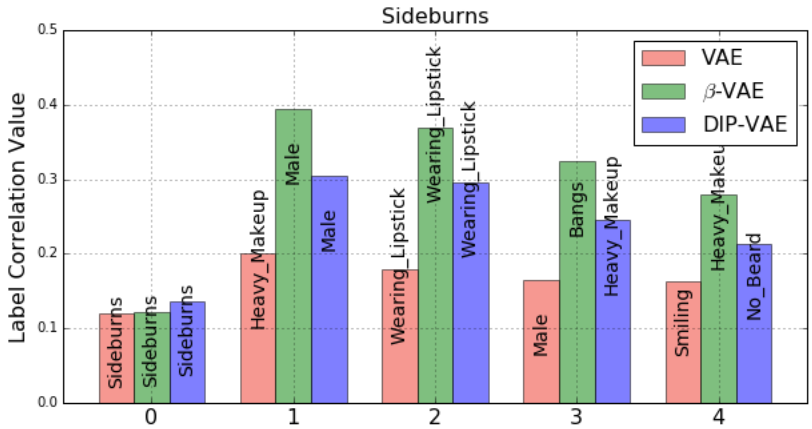

Oval_Face
Label Correlation Value
VAE
β-VAE
DIP-VAE
Oval_Face
Oval_Face
Oval_Face
Smiling
Smiling
Heavy_Makeup
Male
High_Cheekbones
Wearing_Lipstick
Wearing_Lipstick
Mouth_Slightly_Open
Attractive
Big_Nose
Wearing_Earrings
Male
0
1
2
3
4

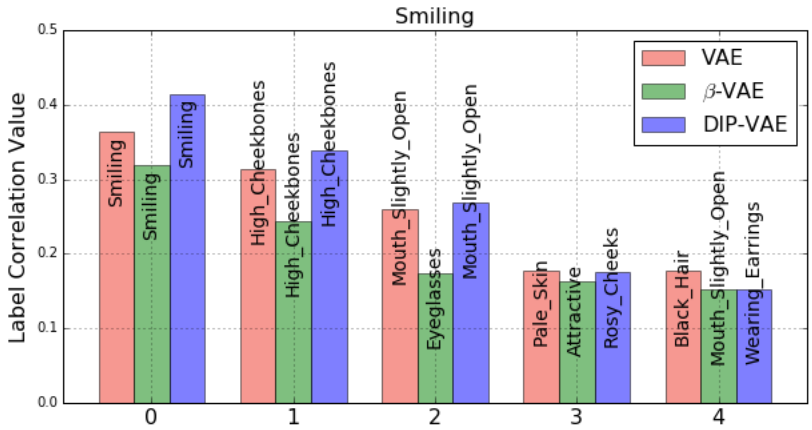

Pale_Skin
Label Correlation Value
VAE
β-VAE
DIP-VAE
Pale_Skin
Pale_Skin
Pale_Skin
Attractive
Smiling
Bangs
Big_Nose
High_Cheekbones
Wearing_Hat
Blond_Hair
Mouth_Slightly_Open
Receding_Hairline
Bangs
Bangs
Male
0
1
2
3
4

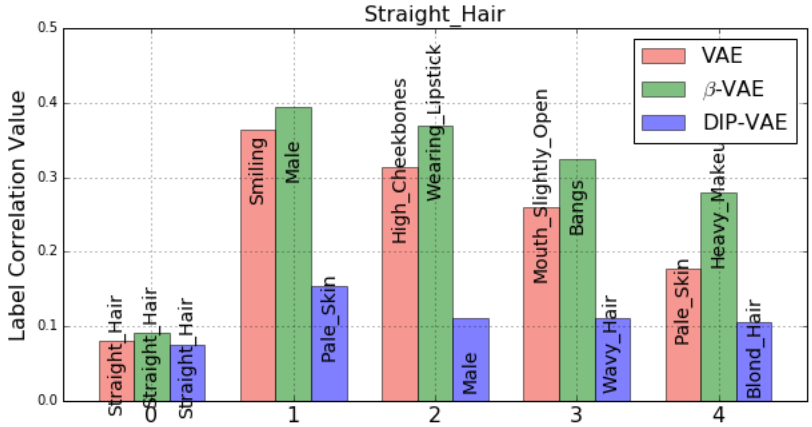

Pale_Skin
Label Correlation Value
0.0
0.1
0.2
0.3
0.4
0.5
VAE
β-VAE
DIP-VAE
Pale_Skin
Pale_Skin
Pale_Skin
Attractive
Smiling
Bangs
Big_Nose
High_Cheekbones
Wearing_Hat
Blond_Hair
Mouth_Slightly_Open
Receding_Hairline
Bangs
Bangs
Male
0
1
2
3
4

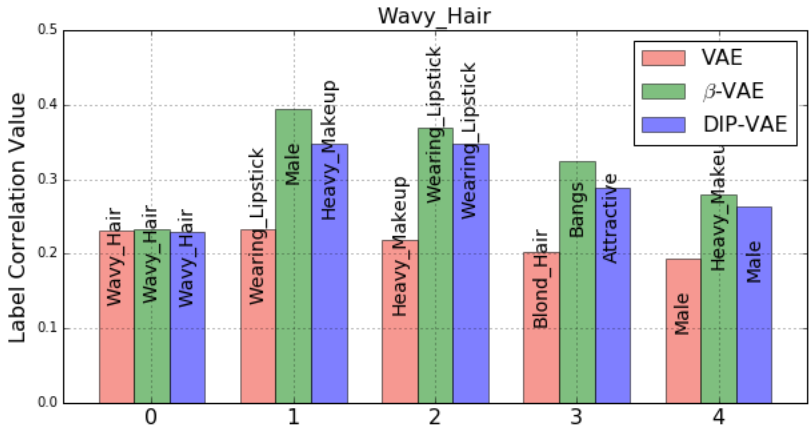

Receding_Hairline
Label Correlation Value
0.0
0.1
0.2
0.3
0.4
0.5
VAE
β-VAE
DIP-VAE
Receding_Hairline
Receding_Hairline
Receding_Hairline
Bangs
Bangs
Attractive
Blond_Hair
Bald
Black_Hair
Black_Hair
Attractive
Blond_Hair
Gray_Hair
Pale_Skin
Young
0
1
2
3
4

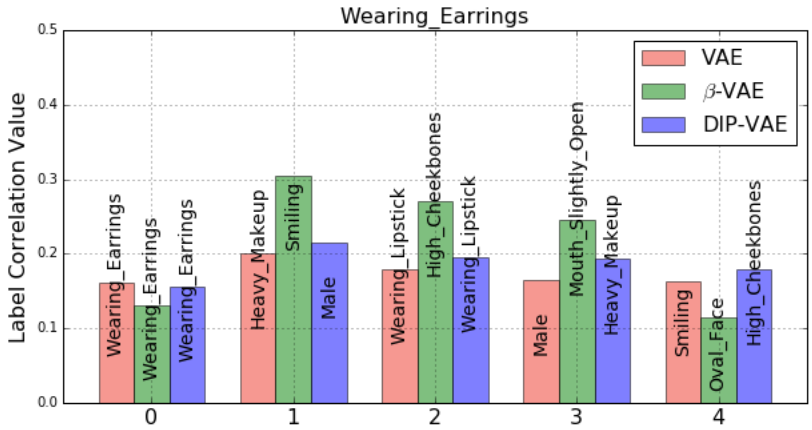

Rosy_Cheeks
Label Correlation Value
0.0
0.1
0.2
0.3
0.4
0.5
VAE
β-VAE
DIP-VAE
Rosy_Cheeks
Rosy_Cheeks
Rosy_Cheeks
Heavy_Makeup
Smiling
Heavy_Makeup
Wearing_Lipstick
High_Cheekbones
Wearing_Lipstick
Male
Eyeglasses
Attractive
Smiling
Attractive
Male
0
1
2
3
4

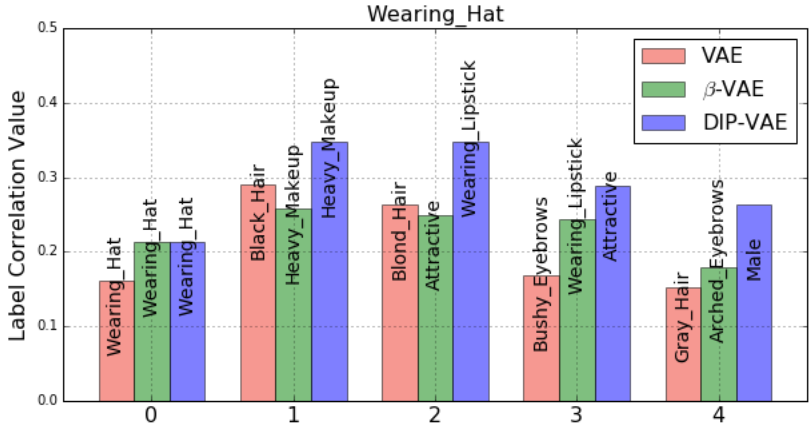

Sideburns
Label Correlation Value
VAE
β-VAE
DIP-VAE
Sideburns
Sideburns
Sideburns
Heavy_Makeup
Male
Male
Wearing_Lipstick
Wearing_Lipstick
Wearing_Lipstick
Male
Bangs
Heavy_Makeup
Smiling
Heavy_Makeup
No_Beard
0
1
2
3
4

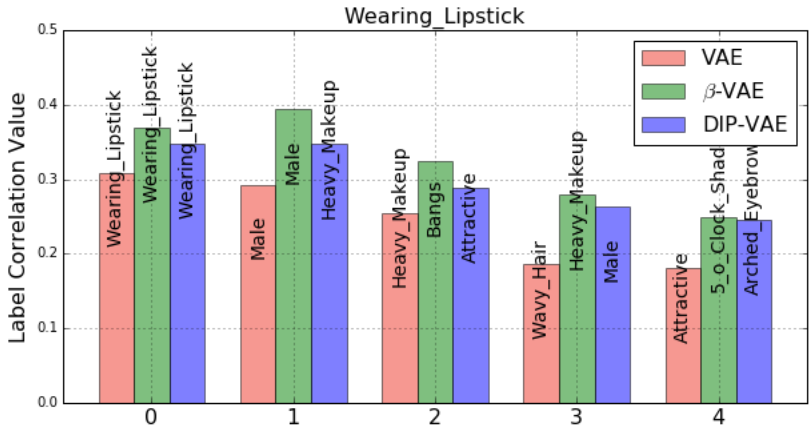

Smiling
Label Correlation Value
VAE
β-VAE
DIP-VAE
Smiling
Smiling
Smiling
High_Cheekbones
High_Cheekbones
High_Cheekbones
Mouth_Slightly_Open
Eyeglasses
Mouth_Slightly_Open
Pale_Skin
Attractive
Rosy_Cheeks
Black_Hair
Mouth_Slightly_Open
Wearing_Earrings
0
1
2
3
4

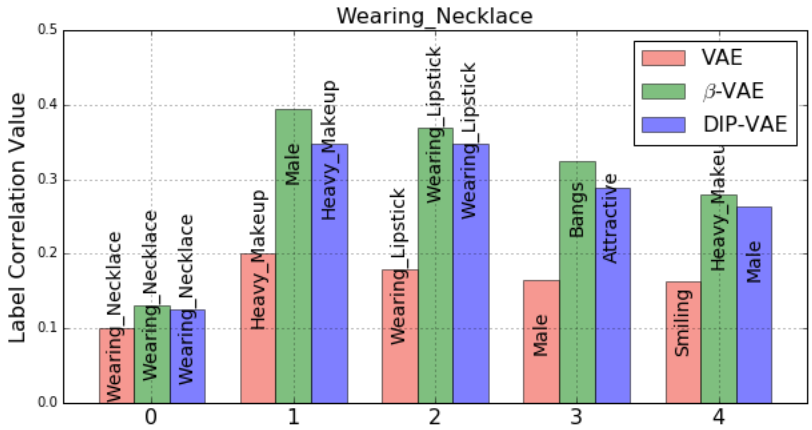

Straight_Hair
Label Correlation Value
VAE
β-VAE
DIP-VAE
Straight_Hair
Straight_Hair
Straight_Hair
Smiling
Male
Pale_Skin
High_Cheekbones
Wearing_Lipstick
Male
Mouth_Slightly_Open
Bangs
Wavy_Hair
Pale_Skin
Heavy_Makeup
Blond_Hair
0
1
2
3
4

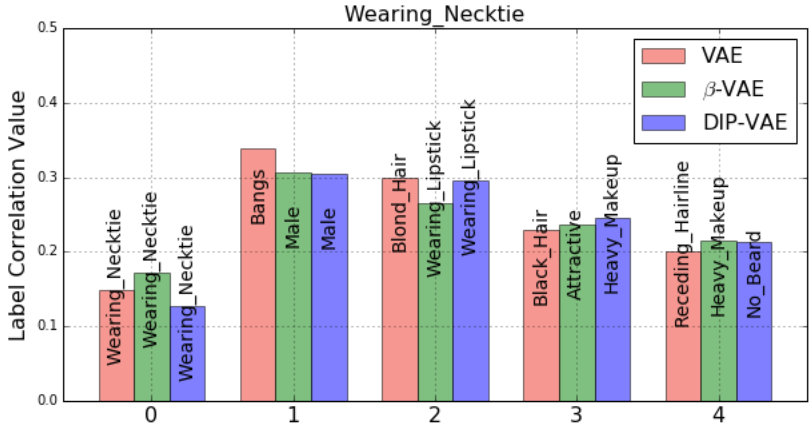

Wavy_Hair
Label Correlation Value
0.0
0.1
0.2
0.3
0.4
0.5
VAE
β-VAE
DIP-VAE
Wavy_Hair
Wavy_Hair
Wavy_Hair
Wearing_Lipstick
Male
Heavy_Makeup
Heavy_Makeup
Wearing_Lipstick
Wearing_Lipstick
Blond_Hair
Bangs
Attractive
Male
Heavy_Makeup
Male
0
1
2
3
4

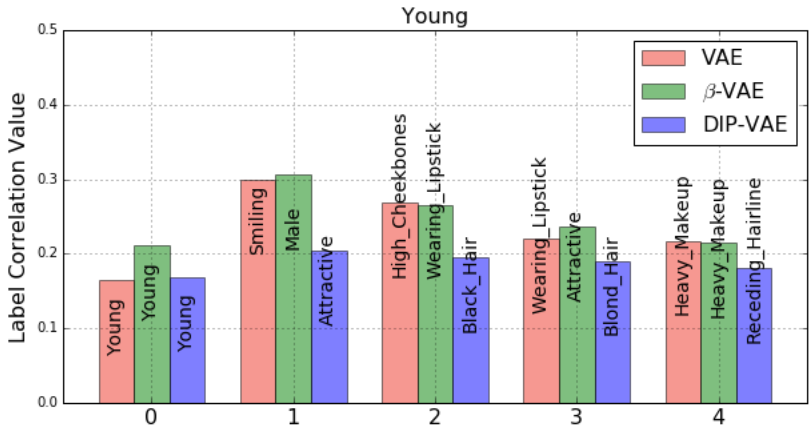

Wearing_Earrings
Label Correlation Value
0.0
0.1
0.2
0.3
0.4
0.5
VAE
β-VAE
DIP-VAE
Wearing_Earrings
Wearing_Earrings
Wearing_Earrings
Heavy_Makeup
Smiling
Male
Wearing_Lipstick
High_Cheekbones
Wearing_Lipstick
Male
Mouth_Slightly_Open
Heavy_Makeup
Smiling
Oval_Face
High_Cheekbones
0
1
2
3
4


Wearing_Hat
Label Correlation Value
0.0
0.1
0.2
0.3
0.4
0.5
VAE
β-VAE
DIP-VAE
Wearing_Hat
Wearing_Hat
Wearing_Hat
Black_Hair
Heavy_Makeup
Heavy_Makeup
Blond_Hair
Attractive
Wearing_Lipstick
Bushy_Eyebrows
Wearing_Lipstick
Attractive
Gray_Hair
Arched_Eyebrows
Male
0
1
2
3
4


Wearing_Lipstick
Label Correlation Value
VAE
β-VAE
DIP-VAE
Wearing_Lipstick
Wearing_Lipstick
Wearing_Lipstick
Male
Male
Heavy_Makeup
Heavy_Makeup
Bangs
Attractive
Wavy_Hair
Heavy_Makeup
Male
Attractive
5_o_Clock_Shad
Arched_Eyebrow
0
1
2
3
4


Wearing_Necklace
Label Correlation Value
VAE
β-VAE
DIP-VAE
Wearing_Necklace
Wearing_Necklace
Wearing_Necklace
Heavy_Makeup
Male
Heavy_Makeup
Wearing_Lipstick
Wearing_Lipstick
Wearing_Lipstick
Male
Bangs
Attractive
Smiling
Heavy_Makeup
Male
0
1
2
3
4


Wearing_Necktie
Label Correlation Value
VAE
β-VAE
DIP-VAE
Wearing_Necktie
Wearing_Necktie
Wearing_Necktie
Bangs
Male
Male
Blond_Hair
Wearing_Lipstick
Wearing_Lipstick
Black_Hair
Attractive
Heavy_Makeup
Receding_Hairline
Heavy_Makeup
No_Beard
0
1
2
3
4



B LABEL CORRELATIONS FOR CELEBA

C LABEL CORRELATIONS FOR DSPRITES

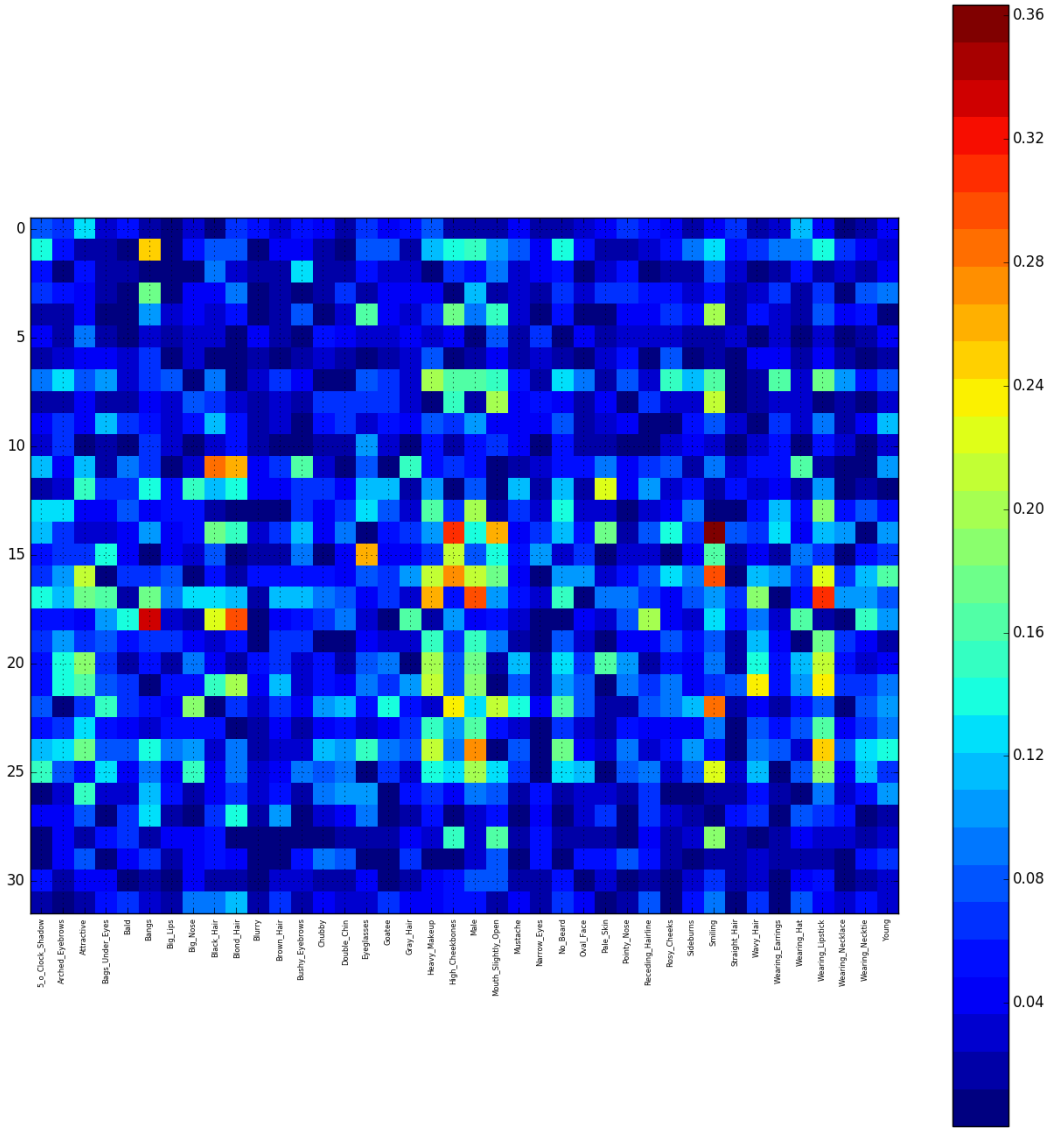


Figure 3: VAE label correlations with attributes for CelebA dataset.

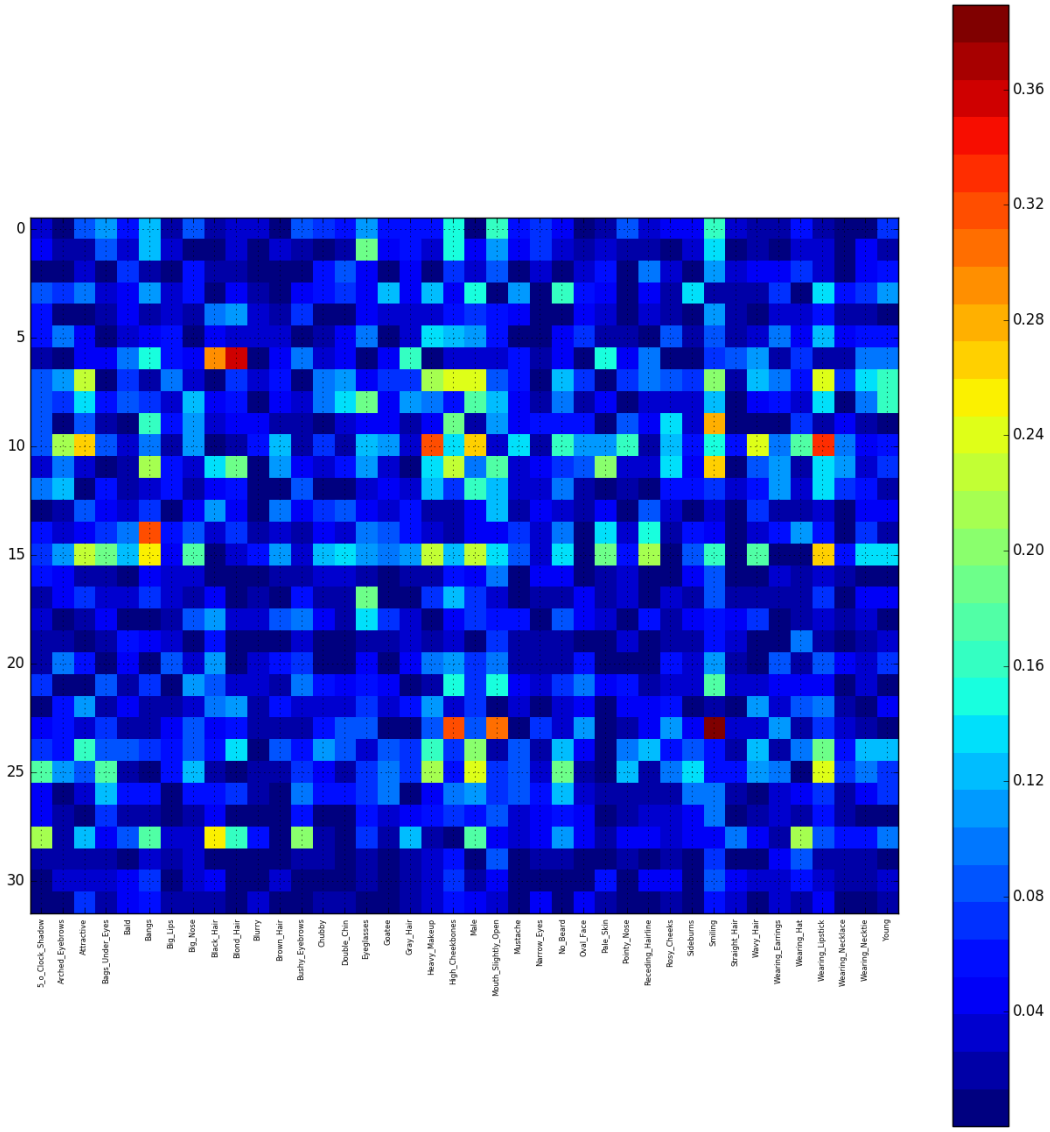


Figure 4: β-VAE ($\beta = 2$) label correlations with attributes for CelebA dataset.

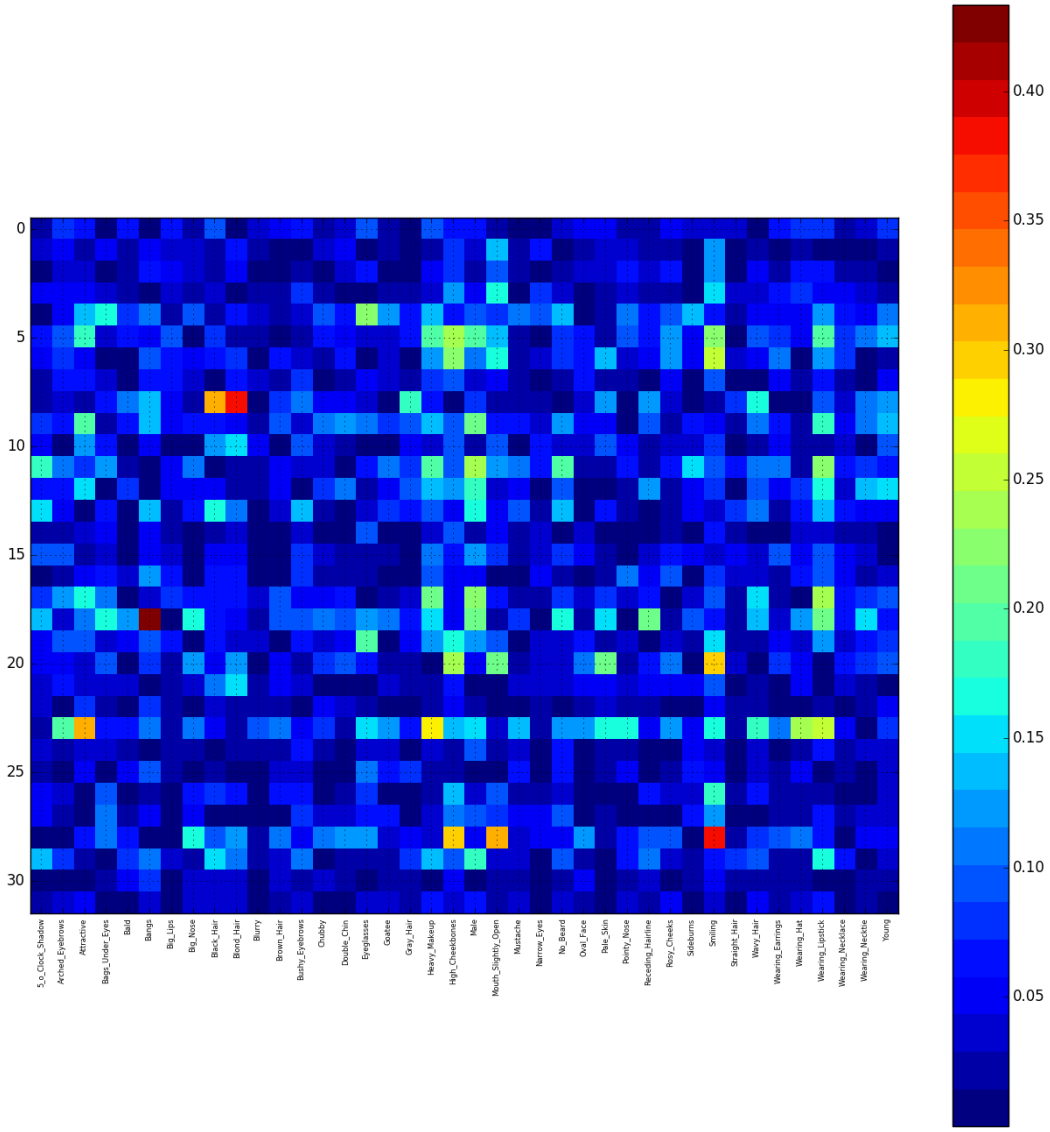


Figure 5: β-VAE ($\beta = 4$) label correlations with attributes for CelebA dataset.

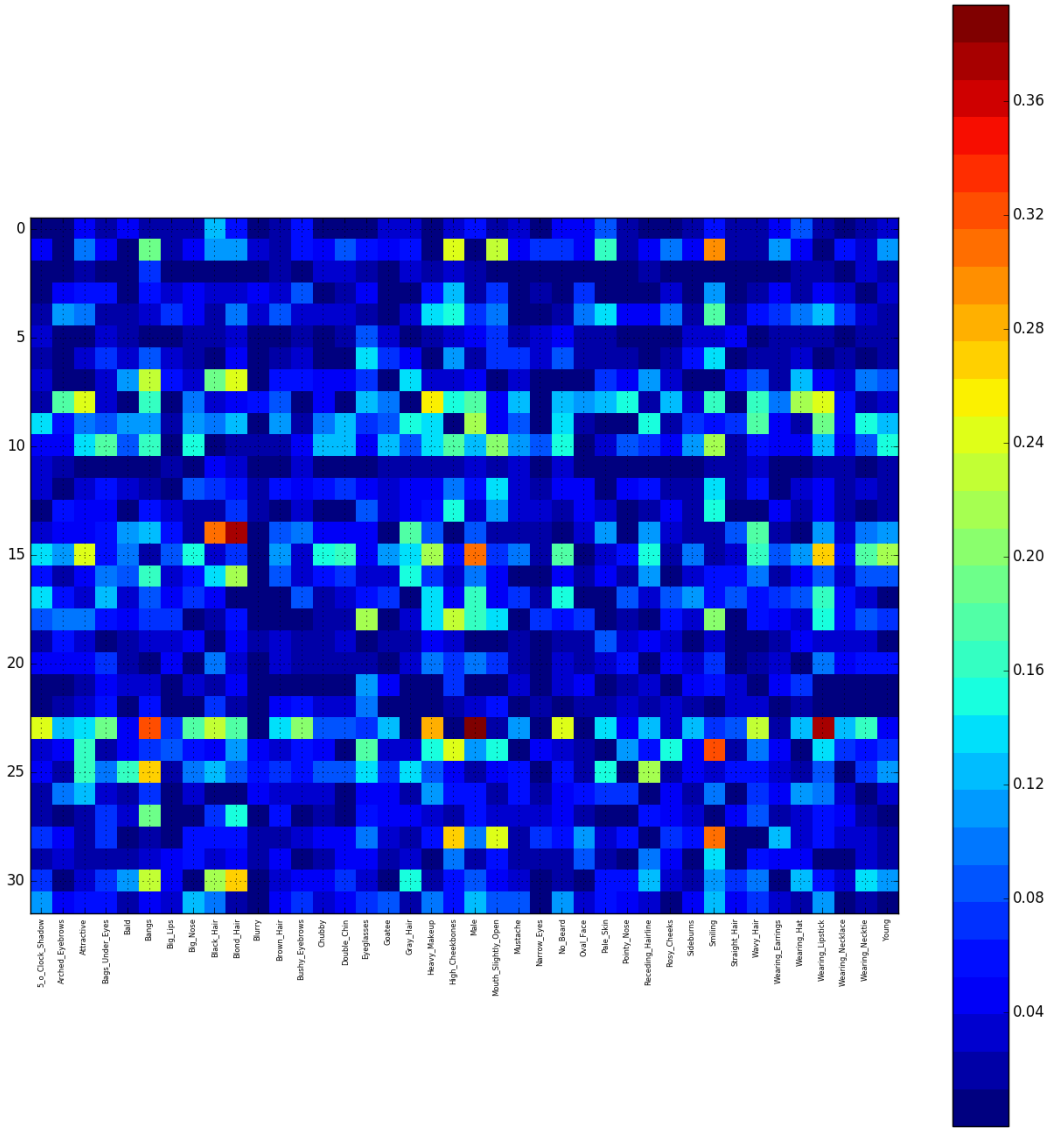


Figure 6: β-VAE ($\beta = 8$) label correlations with attributes for CelebA dataset.

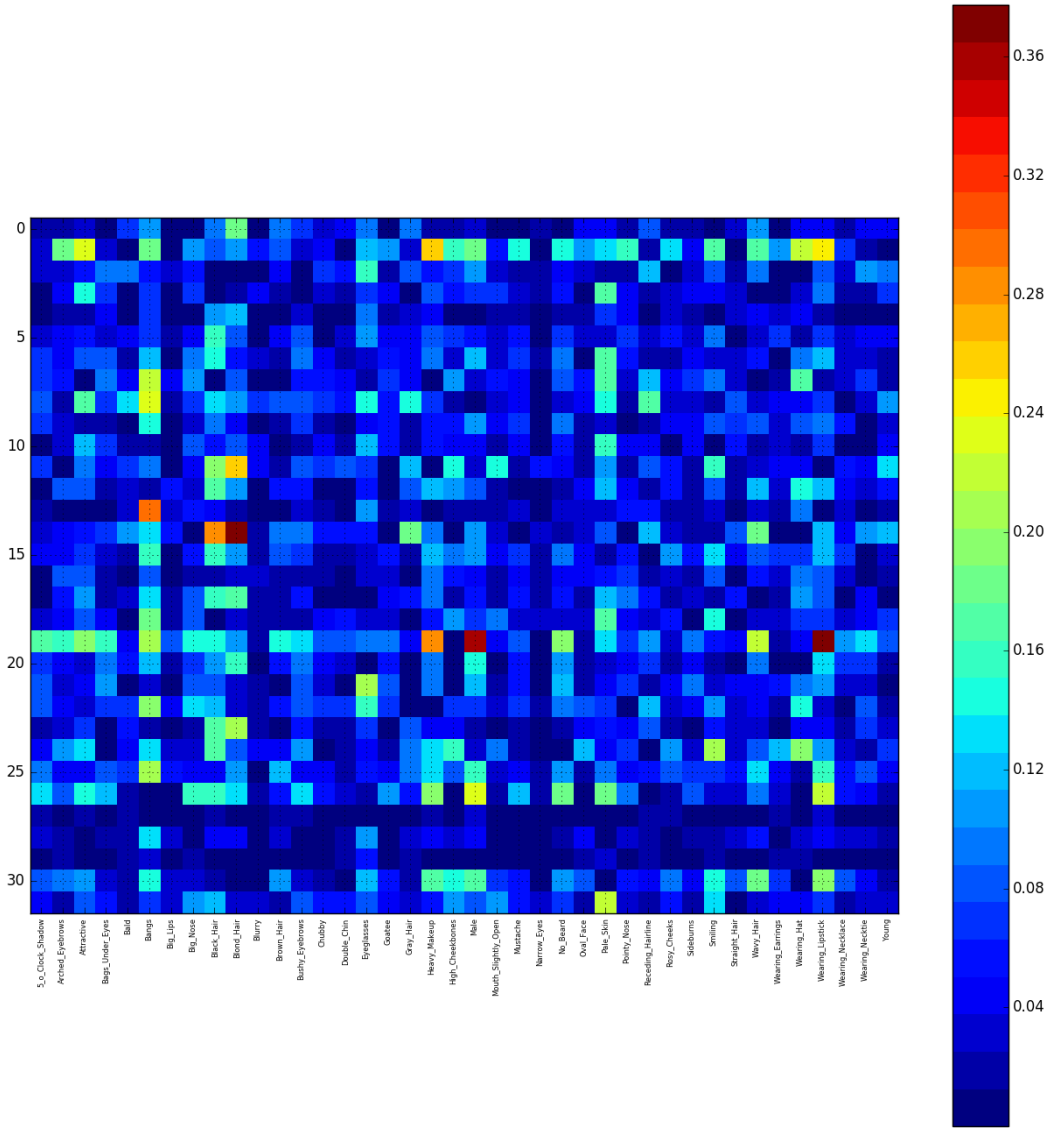


Figure 7: β-VAE ($\beta = 64$) label correlations with attributes for CelebA dataset.

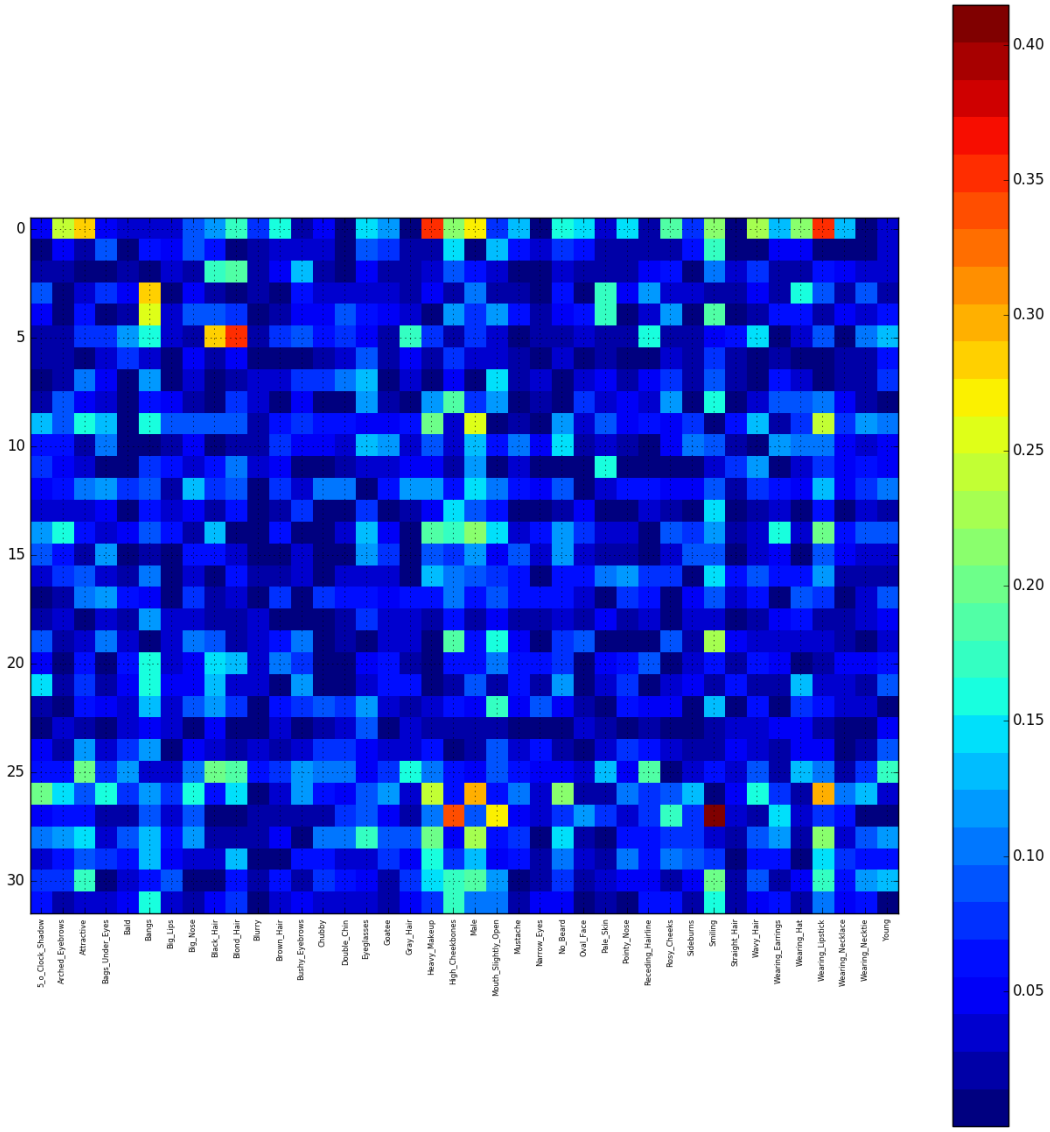


Figure 8: DIP-VAE label correlations with attributes for CelebA dataset.

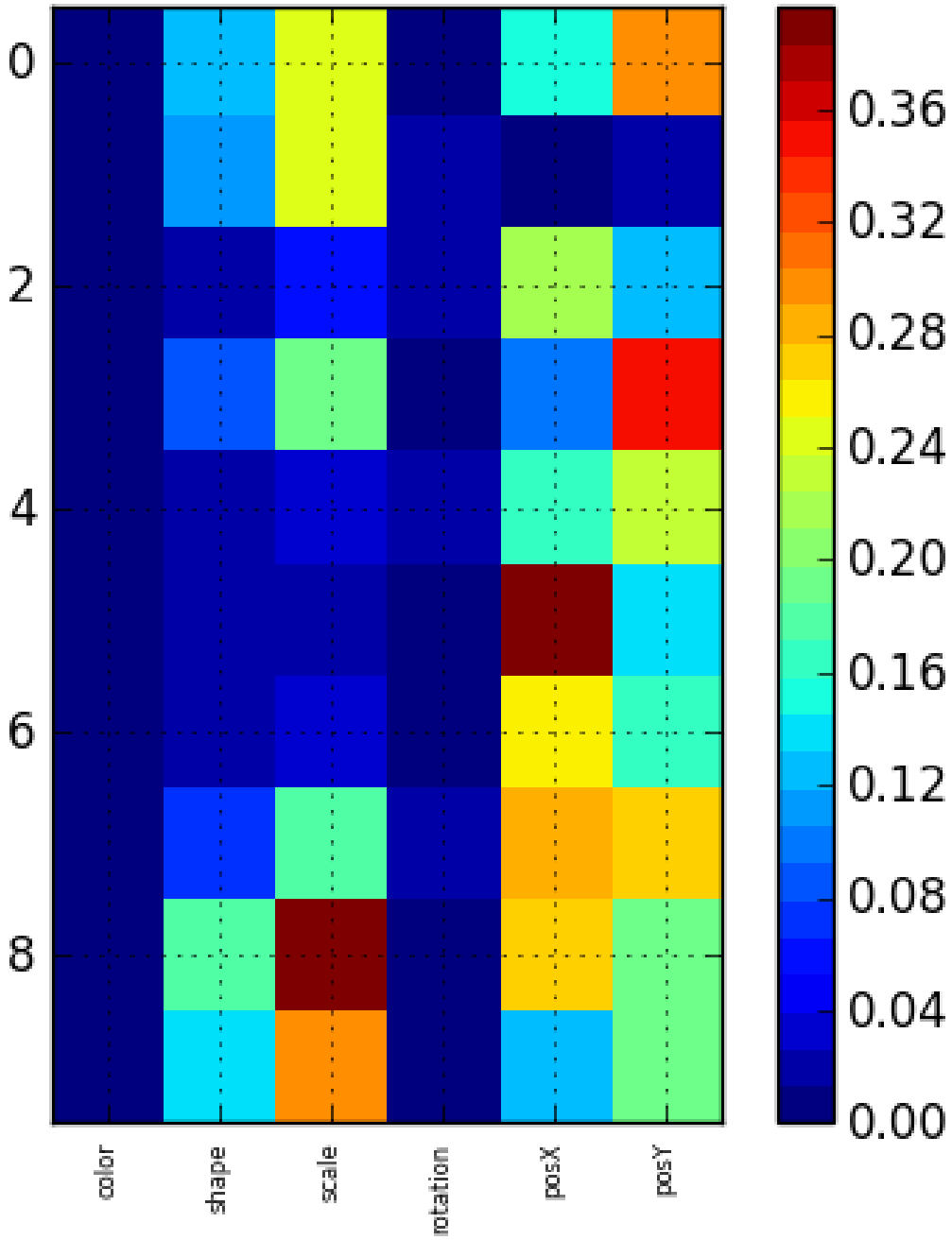


Figure 9: VAE label correlations with attributes for dsprites dataset.

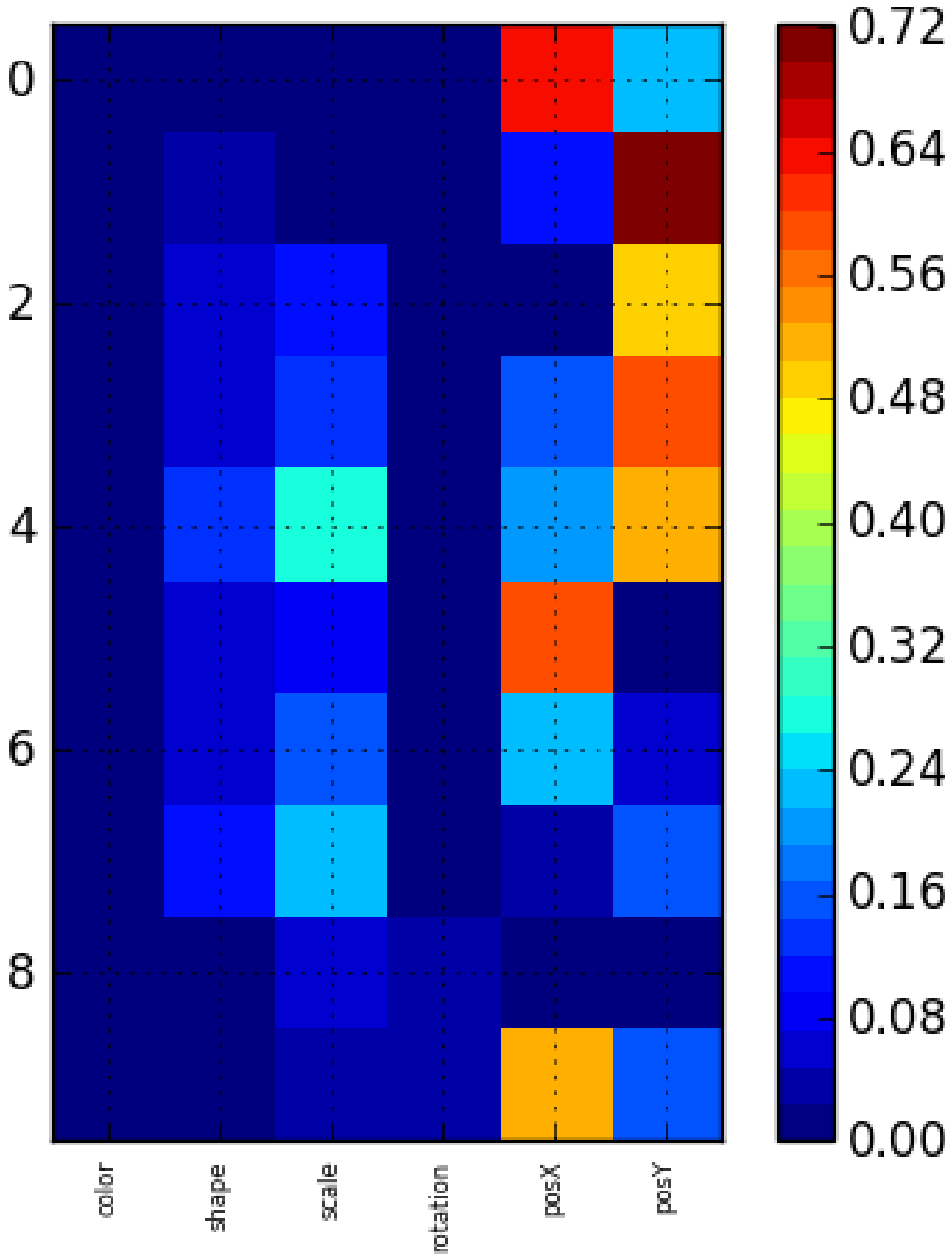


Figure 10: β-VAE ($\beta = 4$) label correlations with attributes for dsprites dataset.

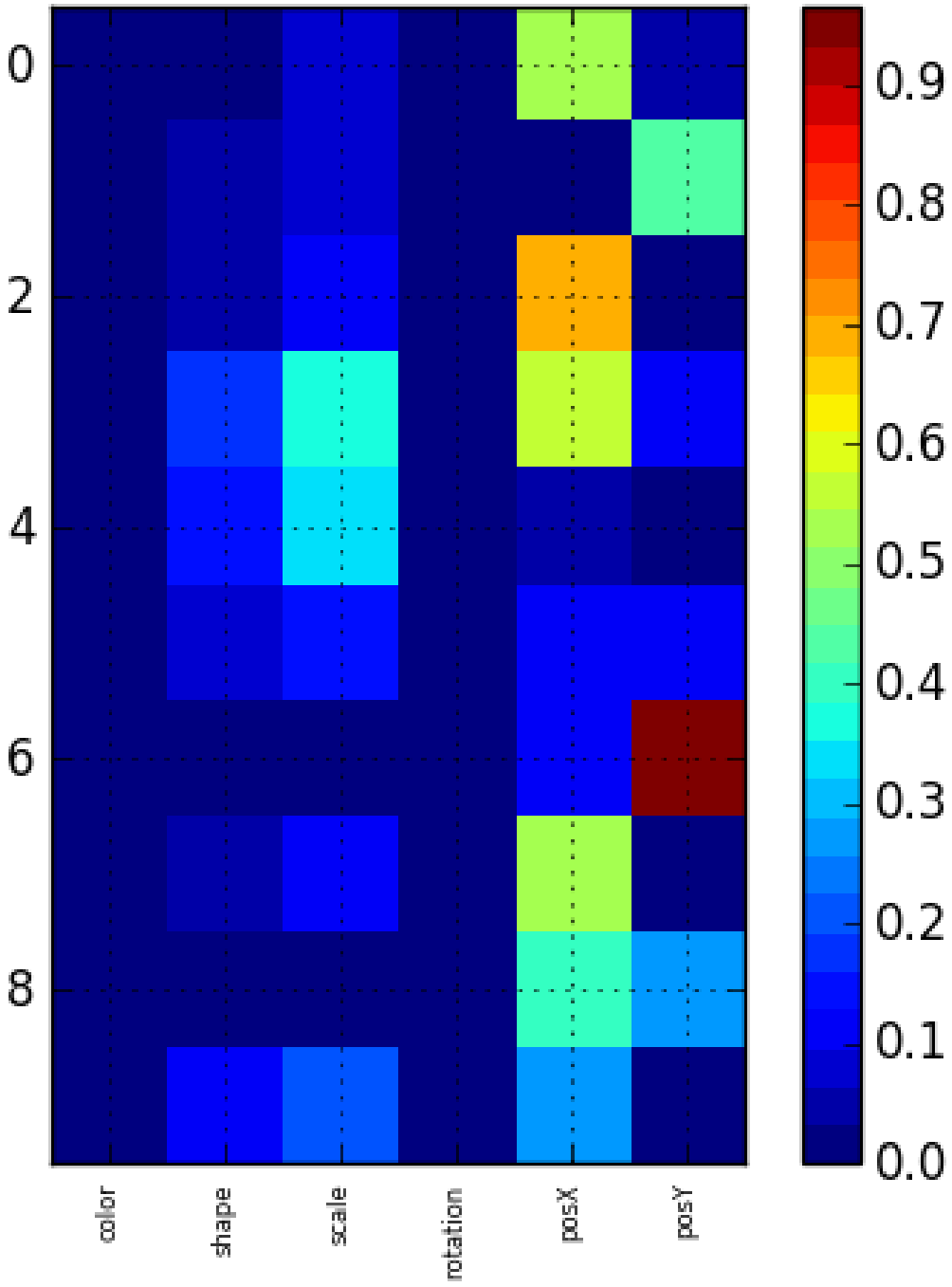


Figure 11: β-VAE ($\beta = 20$) label correlations with attributes for dsprites dataset.

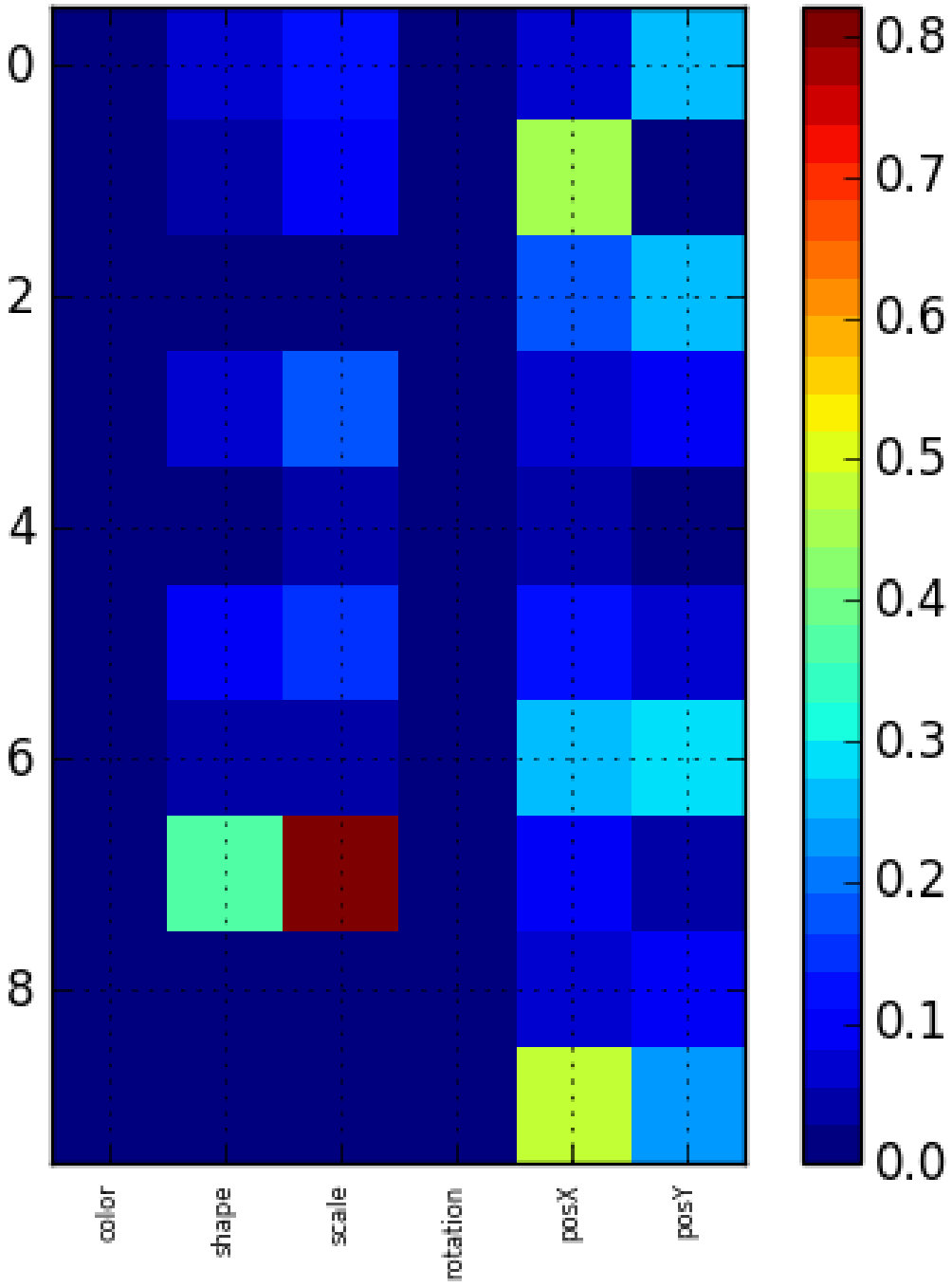


Figure 12: DIP-VAE label correlations with attributes for dsprites dataset.