GROUP-DISENTANGLED REPRESENTATION LEARNING WITH WEAKLY-SUPERVISED REGULARIZATION

Anonymous authors
Paper under double-blind review

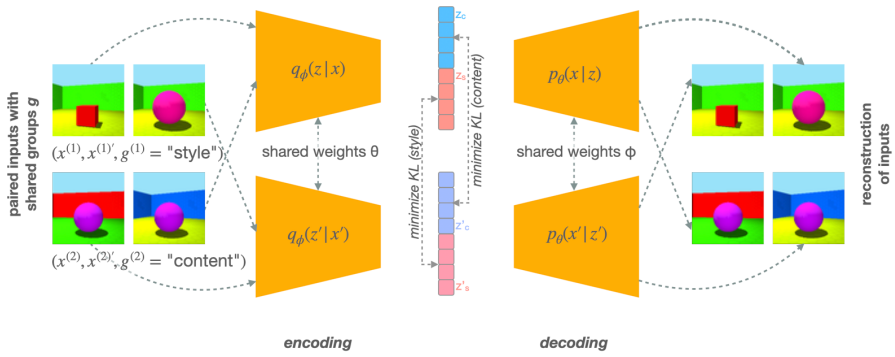
ABSTRACT

Learning interpretable and human-controllable representations that uncover factors of variation in data remains an ongoing key challenge in representation learning. We investigate learning group-disentangled representations for groups of factors with weak supervision. Existing techniques to address this challenge merely constrain the approximate posterior by averaging over observations of a shared group. As a result, observations with a common set of variations are encoded to distinct latent representations, reducing their capacity to disentangle and generalize to downstream tasks. In contrast to previous works, we propose GroupVAE, a simple yet effective Kullback-Leibler (KL) divergence-based regularization across shared latent representations to enforce consistent and disentangled representations. We conduct a thorough evaluation and demonstrate that our GroupVAE significantly improves group disentanglement. Further, we demonstrate that learning group-disentangled representations improve upon downstream tasks, including fair classification and 3D shape-related tasks such as reconstruction, classification, and transfer learning, and is competitive to supervised methods.

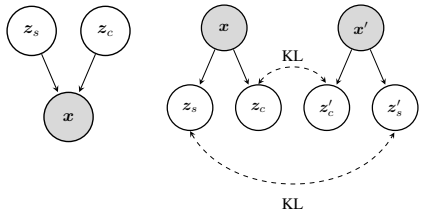
1 INTRODUCTION

Decomposing data into disjoint independent factors of variations, i.e., learning disentangled representations, is essential for interpretable and controllable machine learning (Bengio et al., 2013). Recent works have shown that disentangled representation is useful for abstract reasoning (van Steenkiste et al., 2019), fairness (Locatello et al., 2019a; Creager et al., 2019), reinforcement learning (Higgins et al., 2017b) and general predictive performance (Locatello et al., 2019b). While there is no consensus on the definition of disentanglement, existing works define it as learning to separate all factors of variation in the data (Bengio et al., 2013). According to this definition, altering a single underlying factor of variation should only affect a single factor in the learned representation. However, works in learning disentangled representations (Higgins et al. (2017a); Chen et al. (2018); Locatello et al. (2019b)) have shown that this setting comes with a trade-off between the precision of the representation and the fidelity of the samples. Therefore, learning precise representations for finer factors, i.e., each factor of variation, may not be practical or desirable. We deviate from this stringent assumption to learn *group-disentangled representations*, in which a group might include several factors of variation that can co-variate. For instance, groups of interest may be content, style, or background. As a result, a change in one component might affect other variables in a group but not on other groups.

We present *GroupVAE*, a Variational Autoencoder (VAE) based framework that leverages weak supervision to learn group-disentangled representations. In particular, we use paired observations that always share a group of factors. Existing group-disentangled approaches (Bouchacourt et al., 2018; Hosoya, 2019) enforce disentangled group representations by using an average or product of approximate group posteriors. However, as group representations are dependent on the observations used for the average or product, observations belonging to the same group may not be encoded to the same latent representations. We address this inconsistency challenge by incorporating a simple but effective regularization based on the Kullback-Leibler (KL) divergence. Our idea builds on maximizing the Evidence Lower Bound (ELBO) of the Variational Autoencoders (VAEs) while minimizing the Kullback-Leibler (KL) divergence between the latent variables that correspond to the group shared by the paired observations.



(a) Model



(b) Generative (c) Inference

Figure 1: **GroupVAE's architecture visualization and graphical model.** We visualize the complete model, including model weights in (a) as well as show the (b) generative and (c) inference parts as graphical models. The model visualization shows two paired inputs, one pair sharing "style" and the other sharing "content". The KL minimization depends on the group g that is shared. For instance, for input $(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{'(1)}, g^{(1)} = \text{"style"})$, GroupVAE objective only minimizes the KL between the style latent variables. Shaded nodes denote observed quantities in (b) and (c), and unshaded nodes represent unobserved (latent) variables. Dotted arrows represent minimizing the KL divergence between variables during inference.

In summary, we make the following contributions:

1. We propose a way of learning disentangled representations from paired observations that employs KL regularization between the corresponding groups of latent variables.
2. We propose Group Mutual Information Gap (group-MIG), a mutual information-based metric for evaluating the effectiveness of group disentanglement methods.
3. Through extensive evaluation, we show that our GroupVAE's effectiveness on a wide range of applications. Our evaluation shows significant improvement for group disentanglement, fair facial attribute classification, and 3D shape-related tasks, including generation, classification, and transfer learning.

2 Background & Notation

Variational Autoencoder (VAE). Consider observations $\boldsymbol{X} = \{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(N)}\}$, $\boldsymbol{x}^{(i)} \in \mathbb{R}^D$ sampled i.i.d. from distribution $p_{\boldsymbol{X}}$ and latent variables $\boldsymbol{z}$. A Variational Autoencoder (VAE) learns the joint distribution $p(\boldsymbol{x}, \boldsymbol{z}) = p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$ where $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ is the likelihood function of observations $\boldsymbol{x}$ given $\boldsymbol{z}$, θ are the model parameters of p and $p(\boldsymbol{z})$ is the prior of the latent variable $\boldsymbol{z}$. VAEs are trained to maximize the evidence lower bound (ELBO) on the log-likelihood $\log p(\boldsymbol{x})$. This objective averaged over the empirical distribution is given as

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left(\mathbb{E}_q[\log p(\boldsymbol{x}^{(i)}|\boldsymbol{z})] - \mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)})||p(\boldsymbol{z})), \right. \tag{1}$$

where $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ denotes the learned approximate posterior, ϕ the variational parameters of q and KL denotes the Kullback-Leibler (KL) divergence. VAEs (Kingma & Welling (2014)) are frequently used for learning disentangled representations and serve as the basis of our approach.

Weakly-supervised group disentanglement. We assume the observations $\mathcal{X} = \{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(N)}\}$ and the data generating process can be described by M distinct groups $G = \{g_1, \dots, g_M\}$. Each

group splits $\mathcal{X}$ into disjoint partitions with arbitrary sizes. Each group consists of non-overlapping sets of factors of variations. For example, images of 3D shapes (Burgess & Kim, 2018)[1] can be described through three groups: *shape*[2], *background*[3] and *view*. Without loss of generality, we define two groups g_C (*content*) and g_S (*style* independent of content) to describe the generative and inference process. We assume having paired observations $(\boldsymbol{x}, \boldsymbol{x}')$ for training in a weakly-supervised setting. Each pair of observations shares the same group, i.e., in our case either content $c \in g_C$ or style $s \in g_S$. During inference, the exact values for content and style are unknown, but only that $(\boldsymbol{x}, \boldsymbol{x}')$ share a certain group is known. For each observation $\boldsymbol{x}$, we define two latent variables: $\boldsymbol{z}_c$ for content and $\boldsymbol{z}_s$ for style. The goal for group-based disentanglement is that the representation for the same group as close to each other to ensure consistency.

3 Learning Group-Disentangled Representations

In the following, we introduce *GroupVAE*, a deep generative model which learns disentangled representations for each group of factors. For simplicity, we limit the formulation of GroupVAE to two groups, content and style, although GroupVAE can be applied to any number of groups. This section first describes the generative and inference model and then introduces our main contributions – the KL regularization and inference scheme. We visualized the generative and inference model in Figures 1b and 1c.

Inference and generative model. Our model uses paired observations $(\boldsymbol{x}, \boldsymbol{x}')$ in a weakly-supervised setting. We sample $\boldsymbol{x}$ from the empirical data distribution $p_{\boldsymbol{X}}$ and conditionally sample $\boldsymbol{x}'$ in an i.i.d. manner, so that $\boldsymbol{x}$ and $\boldsymbol{x}'$ belong to the same group g, i.e.,

$$\boldsymbol{x} \sim p_{\boldsymbol{X}}(\boldsymbol{x}); \quad \boldsymbol{x}' \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}'|\boldsymbol{x}, g). \tag{2}$$

Given $\boldsymbol{x}$, we define two latent variables, $\boldsymbol{z}_c$ as content and $\boldsymbol{z}_s$ as style variables. The data is explained by the generative process:

$$p(\boldsymbol{z}_c) = \mathcal{N}(\boldsymbol{z}_c; 0, \boldsymbol{I}); \quad p(\boldsymbol{z}_s) = \mathcal{N}(\boldsymbol{z}_s; 0, \boldsymbol{I}); \quad p(\boldsymbol{x}|\boldsymbol{z}) = p_\theta(\boldsymbol{x}|\boldsymbol{z}_c, \boldsymbol{z}_s) = f_\theta(\boldsymbol{x}; \boldsymbol{z}_c, \boldsymbol{z}_s). \tag{3}$$

Both $\boldsymbol{z}_c$ and $\boldsymbol{z}_s$ are assumed to be independent of each other and are sampled from a Normal distribution with zero mean and diagonal unit variance. f_θ is a suitable likelihood function[4] which is parameterized by a deep neural network. The generative model shown in Figure 1b is also known as the decoding part seen in Figure 1a.

To perform inference, we approximate the true posterior $p(\boldsymbol{z}|\boldsymbol{x})$ with the factorized approximate posterior $q_\phi(\boldsymbol{z}|\boldsymbol{x}) = q_\phi(\boldsymbol{z}_c|\boldsymbol{x}) \cdot q_\phi(\boldsymbol{z}_s|\boldsymbol{x})$ that uses a neural network to amortize the the variational parameters. We specify the inference model as

$$q_\phi(\boldsymbol{z}_c|\boldsymbol{x}) = \mathcal{N}\big(\mu_{\phi,c}(\boldsymbol{x}), \text{diag}(\sigma^2_{\phi,c}(\boldsymbol{x}))\big); \quad q_\phi(\boldsymbol{z}_s|\boldsymbol{x}) = \mathcal{N}\big(\mu_{\phi,s}(\boldsymbol{x}), \text{diag}(\sigma^2_{\phi,s}(\boldsymbol{x}))\big), \tag{4}$$

where both approximate posteriors are assume to be a factorized Normal distributions with mean μ_ϕ and diagonal covariance $\text{diag}(\sigma^2_\phi(\boldsymbol{x}))\big)$. The inference model is visualized as a graphical model in Figure 1c and as the encoding part in Figure 1a. The generative and inference models visualized in Figure 1 apply to $\boldsymbol{x}'$ as well.

VAE objective for paired observation. Given paired observations $(\boldsymbol{x}, \boldsymbol{x}')$, the VAE framework maximizes the Evidence Lower Bound (ELBO)

$$\begin{aligned} \text{ELBO} &= \underbrace{\mathbb{E}_q[\log p(\boldsymbol{x}|\boldsymbol{z}) + \log p(\boldsymbol{x}'|\boldsymbol{z}')]}_{\text{reconstruction losses}} - \underbrace{\text{KL}(q(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z}))}_{\substack{\text{KL between approximate} \\ \text{posterior and prior of } \boldsymbol{z}}} - \underbrace{\text{KL}(q(\boldsymbol{z}'|\boldsymbol{x}')||p(\boldsymbol{z}'))}_{\substack{\text{KL between approximate} \\ \text{posterior and prior of } \boldsymbol{z}'}} \\ &= -\mathcal{L}_{\text{pairedVAE}}, \end{aligned} \tag{5}$$

[1] Samples are shown in Figure 1.

[2] The group *shape* contains factors such as shape category, shape size and shape color.

[3] The group *background* contains factors such as floor color, wall color.

[4] Suitable likelihood functions are, e.g., a Bernoulli likelihood for binary values or a Gaussian likelihood for continuous values.

which consists of the reconstruction losses of the observations $\boldsymbol{x}$ and $\boldsymbol{x}'$ (first two terms) and KL divergence between approximate posterior q and prior p of the latent variables $\boldsymbol{z}$ and $\boldsymbol{z}'$ (third and fourth term). This is a straightforward application of the original ELBO in Equation (1) to two sets of observations, $\boldsymbol{x}$ and $\boldsymbol{x}'$.

KL regularization for group similarity. Rather than defining an average representation for groups as in (Bouchacourt et al., 2018; Hosoya, 2019), we propose to enforce consistency between the latent variables by minimizing KL divergence between the latent variables $\boldsymbol{z}_{=g}$ and $\boldsymbol{z}'_{=g}$. Here, g denotes the group shared between observations $\boldsymbol{x}$ and $\boldsymbol{x}'$. $\boldsymbol{z}_{=g}$ and $\boldsymbol{z}'_{=g}$ denote the corresponding group variable, e.g., if $\boldsymbol{x}$ and $\boldsymbol{x}'$ share group c then the corresponding latent variables are $\boldsymbol{z}_c$ and $\boldsymbol{z}'_c$. Given paired observations $\boldsymbol{x}, \boldsymbol{x}'$ from the same group g, our objective is to minimize

$$\mathcal{L}_{\text{KLreg}} = \text{KL}\big(q(\boldsymbol{z}_{=g}|\boldsymbol{x})||q(\boldsymbol{z}'_{=g}|\boldsymbol{x}')\big). \tag{6}$$

The KL divergence has analytical solutions for Gaussian and Categorical approximate posteriors and is unaffected by the number of shared observations. The analytical solutions can be found in Appendix A.2.

GroupVAE objective and inference. Given a paired observation $(\boldsymbol{x}, \boldsymbol{x}')$ in the sharing group g, we combine the ELBO in Equation (5) and our proposed KL regularization in Equation (6). Our proposed model, *GroupVAE*, has the following minimization objective

$$\mathcal{L}_{\text{GroupVAE}} = \mathcal{L}_{\text{pairedVAE}} + \gamma \mathcal{L}_{\text{KLreg}}, \tag{7}$$

where we treat the degree of regularization γ as a hyperparameter. We propose an alternating inference strategy to encourage variation in both of the latent variables. If we only utilize observations that belong to one group, e.g., paired observations that always share content, we can obtain a trivial solution for the content latent variable by encoding constant latent variables. We overcome this collapse by alternating the group that the observations belong to during training. In particular, during inference we randomly sample a group $g \in \{C, S\}$ and the paired observation $(\boldsymbol{x}, \boldsymbol{x}')$ according to group g. We then minimize the KL divergence of the corresponding latent variable. The inference's pseudo code is shown in Algorithm 1.

Algorithm 1 GroupVAE Inference

while training() **do**
  $g^{(1)}, \dots, g^{(n)} \leftarrow$ getRandomGroups()
  $\boldsymbol{X} \leftarrow$ getMiniBatch()
  $\boldsymbol{X}' \leftarrow$ getPairedObservation($\boldsymbol{X}, g^{(1)}, \dots, g^{(n)}$)
  # encode $\boldsymbol{x}^{(i)}$
  $\forall \boldsymbol{x}^{(i)} \in \boldsymbol{X} : \boldsymbol{z} = (\boldsymbol{z}_c^{(i)}, \boldsymbol{z}_s^{(i)}) \sim q(\boldsymbol{z}_c^{(i)}, \boldsymbol{z}_s^{(i)}|\boldsymbol{x}^{(i)})$
  # encode $\boldsymbol{x}^{(i)\prime}$
  $\forall \boldsymbol{x}^{(i)\prime} \in \boldsymbol{X}' : \boldsymbol{z}' = (\boldsymbol{z}_c^{(i)\prime}, \boldsymbol{z}_s^{(i)\prime}) \sim q(\boldsymbol{z}_c^{(i)\prime}, \boldsymbol{z}_s^{(i)\prime}|\boldsymbol{x}^{(i)\prime})$
  # calculate loss according to Equation (7)
  $\mathcal{L} \leftarrow \sum_i \mathcal{L}_{\text{GroupVAE}}(\boldsymbol{x}^{(i)}, \boldsymbol{x}'^{(i)}, \boldsymbol{z}^{(i)}, \boldsymbol{z}'^{(i)}, g^{(i)})$
  # update gradient $\boldsymbol{g}$ and parameters (θ, ϕ)
  $(\boldsymbol{g}_\theta, \boldsymbol{g}_\phi) \leftarrow (\frac{\partial \mathcal{L}}{\partial \theta}, \frac{\partial \mathcal{L}}{\partial \phi})$
  $(\theta, \phi) \leftarrow (\theta, \phi) + \alpha(\boldsymbol{g}_\theta, \boldsymbol{g}_\phi)$
end while

3.1 Related Work

Unsupervised learning of disentangled representations. Various regularization methods for unsupervised disentangled representation learning have been presented in existing works (Higgins et al., 2017a; Kim & Mnih, 2018; Chen et al., 2018). Even though unsupervised methods have shown promising results to learn disentangled representations, Locatello et al. (2019b) showed in a rigorous study that it is impossible to disentangle factors of variations without any supervision or inductive bias. Since then, there has been a shift towards weakly-supervised disentanglement learning. Our work follows this stream of works and focuses on the weakly-supervised regime instead of an unsupervised one.

Weakly-supervised learning of disentangled representations. Shu et al. (2020) investigated different types of weak supervision and provided a theoretical framework to evaluate disentangled representations. Locatello et al. (2020) proposed to disentangle groups of variations with only knowing the number of shared groups which can be considered as a complementary component to our method. Similar to our method, both these works follow a weakly-supervised setup. However, both approaches focus on the disentanglement of fine-grained factors, whereas our focus is to disentangle groups. Before the concept of paired observations was coined by Shu et al. (2020) as "match pairing", it was already used for geometry and appearance disentanglement (Kossaifi et al., 2018; Tran

et al., 2019) and group-based disentanglement (Bouchacourt et al., 2018; Hosoya, 2019). Closest to our work is MLVAE (Bouchacourt et al., 2018) and GVAE (Hosoya, 2019). For group-disentangled representations, MLVAE uses a product of approximate posteriors, whereas GVAE uses an empirical average of the parameters of the approximate posteriors. A thorough analysis of both works is in Appendix B. In contrast, we employ a simple and effective KL regularization that has no dependency on the batch size.

Alignment between factors of variations and learned representations. Closely related to our work and group-based disentanglement concepts are studies that learn specific latent variables corresponding to one or several factors of variations (or labels). Dupont (2018) used both continuous and discrete latent variables to improve unsupervised disentanglement of mixed-type latent factors. Creager et al. (2019) proposed to minimize the mutual information between the sensitive latent variable and sensitive labels. Similarly, Klys et al. (2018) proposed to minimize Mutual Information (MI) between the latent variable and a conditional subspace. Both works (Creager et al., 2019; Klys et al., 2018) require either supervision, sensitive labels, or conditions to estimate the mutual information, whereas we only use weak supervision for learning disentangled group representations. Concurrent to our work, Sinha & Dieng (2021) proposed to use a KL regularization for learning a VAE with representation that is consistent with augmented data. While Sinha & Dieng (2021) use the KL regularization to enforce the encoding to be consistent with changes in the input, our goal is to split the representation into subspaces that correspond to the different groups of variations.

4 Evaluation

Here, we evaluate our GroupVAE and compare it to existing approaches. We show that our approach outperforms existing approaches for group-disentanglement and disentanglement on existing disentanglement benchmarks. Within the context of evaluating group disentanglement, we propose a MI-based evaluation metric to assess the degree of group disentanglement. Further, we demonstrate that our approach is generic and can be applied to various applications, including fair classification and 3D shape-related tasks (reconstruction, classification, and transfer learning).

4.1 Weakly-supervised group-disentanglement

Experimental settings. We used three standard datasets on disentangled representation learning: 3D Cars (Reed et al., 2014), 3D Shapes (Burgess & Kim, 2018) and dSprites (Matthey et al., 2017). Despite the fact that these image datasets are synthetic, disentangling the factors of variation remains a difficult and unresolved task (Locatello et al., 2019b; 2020). We use Mutual Information Gap (MIG) (Chen et al., 2018) and our proposed metric group-MIG for quantitative evaluation different approaches. We compare our model, GroupVAE, to unsupervised methods (β-VAE (Higgins et al., 2017a) and FactorVAE (Kim & Mnih, 2018)) as well as weakly-supervised methods (AdaGVAE (Locatello et al., 2019b), MLVAE (Bouchacourt et al., 2018), and GVAE (Hosoya, 2019)). For all methods, we ran a hyperparameter sweep varying regularization strength for five different seeds. We report the median group-MIG and MIG.

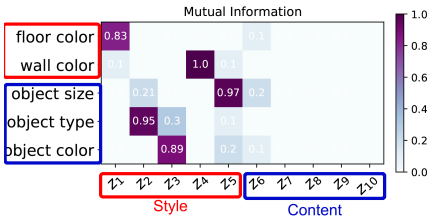


Figure 2: **Example of failed content-style disentanglement with high MIG.** The heatmap shows the MI of each pair of factors and latent dimensions. Although content and style have not been separated in the corresponding latent dimensions, the MIG is still very high ($= 0.76$). In contrast, group-MIG considers where the groups are captured, and thus, the group-MIG is much lower ($= 0.34$).

group-MIG for evaluating group disentanglement. The Mutual Information Gap (MIG) (Chen et al., 2018) is a commonly used evaluation metric for disentanglement. This metric measures the normalized difference between the latent variable dimensions with highest and second-highest MI values. The higher the MIG, the greater the degree of disentanglement is. However, MIG can still be high if the style latent variable disentangles all factors of variation whereas the content variable collapse to a constant value. An example of a failure in group disentanglement is shown in Figure 2. Therefore, we introduce group-MIG, a metric based on MIG, which addresses this

Table 1: **Quantitative disentanglement results.** We report median group-MIG and median MIG over five hyperparameter sweeps of different seeds (*higher is better*). Since the unsupervised approaches and AdaGVAE do not learn group disentangled representations, we cannot report group-MIG for these groups and denote it with $-$. We highlight in **bold** the best results.

Type	Model	3D Cars		3D Shapes		dSprites	
		group-MIG	MIG	group-MIG	MIG	group-MIG	MIG
unsup.	β-VAE	–	0.08	–	0.22	–	0.10
unsup.	FactorVAE	–	0.10	–	0.27	–	0.14
weakly-sup.	AdaGVAE	–	0.15	–	**0.56**	–	0.26
weakly-sup.	MLVAE	0.24	0.07	0.47	0.32	0.11	0.22
weakly-sup.	GVAE	0.27	0.08	0.45	0.31	0.14	0.21
weakly-sup.	GroupVAE (ours)	**0.48**	**0.18**	**0.60**	0.31	**0.54**	**0.27**

issue and quantitatively estimates the mutual information between groups and corresponding latent variables. We define group-MIG as

$$\frac{1}{m}\sum_{i=1}^{m}\frac{1}{H(g_i)}\left|\max I(z_{=g_i};g_i)-\max_{j\neq i} I(z_{\neq g_j};g_i)\right|, \tag{8}$$

where m is the number of groups, g_i is the ground truth group, and $I(\boldsymbol{z};g_i)$ is an empirical estimate of the MI between continuous variable $\boldsymbol{z}$ and g_i. The values of group-MIG is small if the group factors are not represented in the corresponding latent vectors, even though the factor is disentangled within the other variables.

Group labeling. We define the following groups based on the fine-grained factors for each dataset:

- dSprite:s $g_C = [\text{shape}, \text{scale}],\ g_S = [\text{orientation}, \text{x-position}, \text{y-position}]$
- 3D Shapes: $g_C = [\text{obj. color}, \text{obj. size}, \text{obj. type}],\ g_S = [\text{floor color}, \text{wall color}, \text{azimuth}]$
- 3D Cars: $g_C = [\text{obj. type}],\ g_S = [\text{elevation}, \text{azimuth}]$

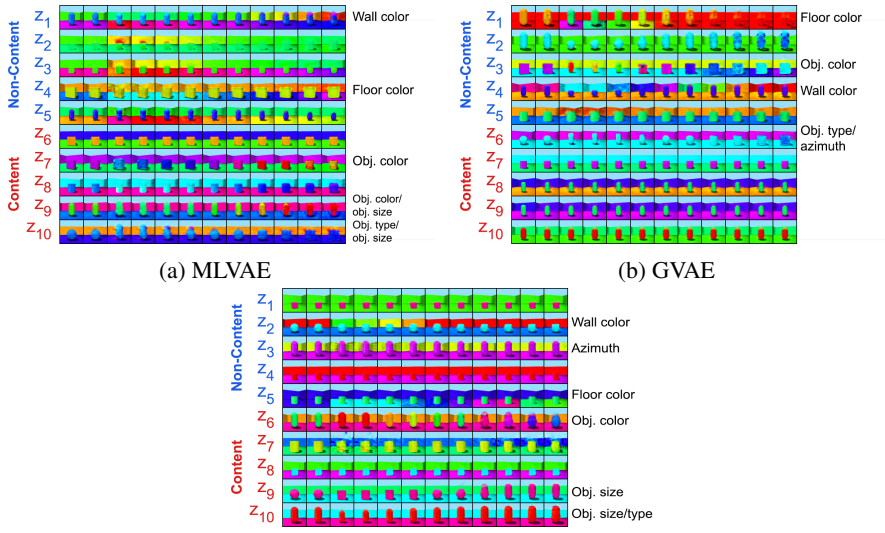
Results. We consistently outperform weakly-supervised disentanglement models w.r.t. median group-MIG over five hyperparameter sweeps of different seeds by *at least 25%* (3D Shapes). Further, we also improve on disentanglement w.r.t. MIG for two out of three datasets (3D Cars, dSprites). In addition, we show interpolation samples of MLVAE, GVAE, and GroupVAE[5] for 3D Shapes in Figure 3. Both MLVAE and GVAE are not able to capture azimuth in the latent representations. Moreover, GVAE encodes almost all factors into the style part and collapses to a constant representation in the content part. The interpolations of GroupVAE show content and style disentanglement, although some factors such as object size and type for 3D Shapes remain entangled. As we assume that factors in a group can co-variate, this result is expected as object size and type are in the same group.

4.2 Application to fair classification

We examine the problem of learning fair representations for classification problems as an application of our method. In particular, we want to learn fair group representation in which members of any (demographic) groups have an equal probability of being assigned to the positive predicted class. Deep learning algorithms have been proven to be biased against specific demographic groups or populations (Mehrabi et al., 2021). It is critical that classification models can produce accurate predictions without discriminating against certain groups in high-stakes and safe-related applications. In this context, we propose to learn fair representations by learning two distinct groups of representations: a predictive representation for evaluating the downstream task and a representation to account for the sensitive factors, e.g., gender- or age-specific attributes. The latter representation is solely utilized for training and not for downstream tasks.

Learning fair representations consist of a two-step optimization scheme. First, we train GroupVAE with pairs of observations sharing either sensitive and non-sensitive attributes. Second, we train a simple MLP for attribute classification using the non-sensitive mean representation. We measure

[5]We selected models with median group-MIG over five hyperparameter sweeps of different seeds.



(a) MLVAE (b) GVAE

(c) GroupVAE (ours)

Figure 3: **Interpolations of 3D Shapes.** We show samples from our model GroupVAE and the baseline models (MLVAE and GVAE) with median group-MIG over five hyperparameter sweeps. For each subplot, we show random inputs (first column), its reconstructions (second column) and reconstruction when interpolating the latent variables (remaining columns) of each latent dimension (row-wise). The factors annotated on the right side are those with a high level of mutual information (MI > 0.25). For all three models z_1 to z_5 is supposed to capture style (non-content) while z_5 to z_{10} is supposed to capture content.

Table 2: **Fair classification results on the test set of dSpriteUnfair and CelebA.** We report test accuracy and Demographic Parity (DP) for each sensitive attribute with an average of five experiments. We report the standard error for all test accuracies, but leave out the standard error for all DP results as they were < 0.002. We highlight in **bold** the best results. The column *Fair learning* refers to whether a model uses any supervision during the fair representation learning phase. For the final classification, all models use full supervision.

			Demographic parity (DP) $\downarrow \in [0, 1]$	
Fair learning	**Model**	Test acc. $\uparrow$	"shape" $\downarrow$	"scale" $\downarrow$
✗	MLP	99.07 $\pm$0.06	0.007	0.008
✗	CNN	99.04 $\pm$0.05	**0.002**	**0.002**
✓ (supervised)	FFVAE	98.60 $\pm$0.12	0.004	0.004
✓ (weakly-superv.)	GroupVAE	**99.18** $\pm$0.08	**0.002**	**0.002**

(a) Results for dSpritesUnfair predicting "x-position".

			Demographic parity (DP) $\downarrow \in [0, 1]$	
Fair learning	**Model**	**Test acc.** $\uparrow$	"Male"	"Young"
✗	MLP	97.89 $\pm$0.01	0.99	0.99
✗	CNN	**98.46** $\pm$0.03	0.95	0.93
✓ (supervised)	FFVAE	97.79 $\pm$0.01	0.04	0.04
✓ (weakly-superv.)	GroupVAE	98.23 $\pm$0.02	**0.01**	**0.02**

(b) Results for CelebA predicting "bald".

			Demographic parity (DP) $\downarrow \in [0, 1]$			
Fair learning	**Model**	**Test acc.** $\uparrow$	'BigNose'	'HeavyMakeup'	'Male'	'WearingLipstick'
✗	MLP	77.24 $\pm$0.29	0.09	0.15	0.06	0.04
✗	CNN	79.90 $\pm$0.06	0.11	0.15	0.03	0.06
✓ (supervised)	FFVAE	97.75 $\pm$0.03	0.03	**0.02**	0.03	0.03
✓ (weakly-superv.)	GroupVAE	**97.88** $\pm$0.01	**0.01**	**0.02**	**0.02**	**0.01**

(c) Results for CelebA predicting "attractive".

classification accuracy and Demographic Parity (DP). DP measures whether the predictive outcome is independent of a sensitive attribute. A completely fair model would attain a DP value of 0.0, whereas a biased model can have a DP up to 1.0. We compare against MLP and CNN baselines,

and FFVAE (Creager et al. (2019)) which learns fair representations by using a supervised loss on the sensitive attributes and a total correlation loss. We used two datasets: dSpritesUnfair (Creager et al. (2019); Träuble et al. (2020)) and CelebA (Liu et al. (2015)). dSpritesUnfair is a modified image dataset based on dSprites with binarized factors of variations and is sampled with shape and x-position being highly correlated. For CelebA, an image dataset of celebrity faces with 40 binary attribute labels, we predict "bald" and "attractive" in two separate experiments. For predicting "bald", we use the attributes "male" and "young" as sensitive attributes whereas we use the attributes "BigNose", "HeavyMakeUp", "Male" and "WearingLipstick" as sensitive attributes for predicting "attractive". We argue that these attributes have a weak correlation but a strong correlation with the predictive attribute. However, several CelebA attributes significantly correlate, making this a difficult dataset for fairness classification. We refer to the Appendix C.2 for the detailed experimental settings.

Results. We report the fair classification results in Table 2. Overall, the results in Table 2 show that weakly-supervised fair representation learning (GroupVAE) outperforms supervised fair representation learning (FFVAE). Further, we either get competitive or even outperform the supervised baselines (MLP, CNN). Surprisingly, when evaluating dSpritesUnfair the demographic parity for all models is relatively low, and the strong correlation between shape and x-position does not seem to affect the classification. The test accuracy and DP of the sensitive attributes of all the competitive models are very close to each other. Nevertheless, among all models, our method achieves the highest test accuracy and lowest DPs. For predicting "bald" in CelebA, even though both MLP and CNN baselines achieve high test accuracy, the DPs shows an extremely biased classification towards gender-specific and male-specific attributes. In contrast, our GroupVAE achieves the lowest DPs but still attain competitive classification accuracy, i.e., second highest test accuracy after the CNN performance. When predicting "attractive", GroupVAE decreases the bias of all sensitive attributes and increases the test accuracy compared to all other models.

4.3 Application to 3D point cloud tasks

In addition to evaluating image datasets, we show experiments on 3D point clouds for reconstruction and classification. We experimented with FoldingNet (Yang et al. (2018)), a deep autoencoder that learns to reconstruct 3D point clouds in an unsupervised way. Unlike VAEs, the FoldingNet autoencoder is deterministic and does not optimize the representation to be a probabilistic distribution. Instead of converting the autoencoder into a VAE, we use a similar approach as Ghosh et al. (2019). We assume the embedding of autoencoder to be Normal distributed with constant variance. Given this assumption, the KL divergence between the corresponding embeddings reduces to a simple L2 regularization, and we can inject noise to regularize the decoding. We evaluate three tasks, 3D point cloud reconstruction, classification, and transfer learning. We measure the Chamfer Distance (CD) and the Earth Mover's Distance (EMD) to assess reconstruction quality and report accuracy to assess classification and transfer learning performance. We compare to FoldingNet (unsupervised) and DGCNN (supervised) (Wang et al. (2019)), a dynamic graph-based classification approach. For assessing the transfer learning capability, we use a linear SVM classifier on the extracted representation. We used two datasets for training: FG3D (Liu et al. (2021)) and ShapeNetV2 (Chang et al. (2015)). FG3D contains 24,730 shapes with annotations of basic categories (Airplane, Car, and Chair) and fine-grained sub-categories. ShapeNetV2 contains 51,127 shapes with annotations of 55 categories. For transfer learning, we also use ModelNet40 (Wu et al. (2015)).

Results. Table 3a shows that weakly-supervised training improves upon 3D point cloud reconstruction for both FG3D and ShapeNetV2. Table 3b shows the classification and transfer results. Our approach GroupFoldingNet improves point cloud classification compared to the original FoldingNet and is competitive with the supervised approach when training with FG3D. We outperform both supervised and unsupervised transfer learning performances when training with FG3D and evaluating ShapeNetV2 and ModelNet40. We are competitive to the supervised approach when training with ShapeNetV2 and evaluating on ModelNet40. In particular, the transfer learning performance with FG3D as the training set highlights the capabilities of weakly-supervised group disentanglement as it can learn 3D point clouds of three classes and transfer it to ShapeNetV2, a large-scale dataset with 55 classes. We also visualize point cloud reconstructions and interpolations of three different classes using our approach in Figure 4. The reconstructions show that our approach is better than

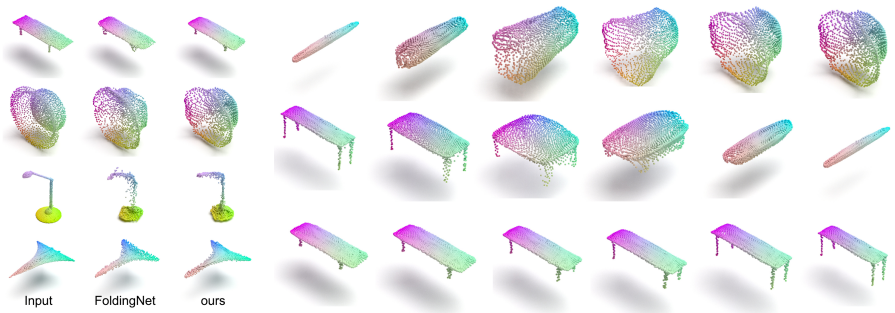
Table 3: **Evaluation of 3D point cloud reconstruction, classification, and transfer learning.** We report Chamfer Distance (CD) and Earth Mover Distance (EMD) for quality of reconstruction and accuracy for classification and transfer learning tasks. Best results without full supervision are highlighted in **bold**.

		Reconstruction ↓			
		FG3D		ShapeNetV2	
Type	**Model**	CD	EMD	CD	EMD
unsupervised	FoldingNet	0.9539	0.9340	2.9867	1.5576
weakly-superv.	GroupFoldingNet (ours)	**0.7519**	**0.8191**	**2.6891**	**1.3009**

(a) Reconstruction results for FG3D and ShapeNetV2.

						Linear SVM ACC ↑	
Type	**Model**	**Training dataset**	**Test dataset**	**#classes**	**Test ACC ↑**	ShapeNetV2	ModelNet40
supervised	DGCNN	FG3D	FG3D	3	99.26	50.53	74.25
unsupervised	FoldingNet	FG3D	FG3D	3	98.27	85.45	80.04
weakly-superv.	ours	FG3D	FG3D	3	**98.57**	**87.24**	**81.39**
supervised	DGCNN	ShapeNetV2	ShapeNetV2	55	94.4	–	90.02
unsupervised	FoldingNet	ShapeNetV2	ShapeNetV2	55	81.51	–	87.40
weakly-superv.	ours	ShapeNetV2	ShapeNetV2	55	**82.62**	–	**89.97**

(b) Classification and transfer learning of representations.



(a) Reconstructions. (b) Interpolations between two different samples.

Figure 4: **Qualitative samples of ShapeNetV2.** We show reconstructions of FoldingNet and our approach in (a) and show interpolations of our approach in (b).

FoldingNet in reconstructing finer details. Further, the interpolations show that our approach can learn an interpretable representation.

5 Conclusion & Discussion

We proposed a simple KL regularization for VAEs to enforce group disentanglement through weak supervision. We empirically showed that our model outperforms existing approaches in group disentanglement. Further, we demonstrated that learning group-disentangled representations outperforms performance on fair image classification and 3D shape-related tasks (reconstruction, classification, and transfer learning) and is even competitive to supervised approaches.

There are several possible directions for future work. In comparison to unsupervised representation learning, weakly-supervised learning, by definition, requires some weak form of supervision. Although we only need knowledge of whether two observations share a specific group, this limits the approach. Further, we require group labels for the entire dataset for training and evaluation. For real-life applications, datasets may not be fully labeled, and performance may suffer under this setting. Future investigation of group disentanglement in a low data or a "semi" weakly-supervised regime can allow group disentanglement learning to transfer to large-scale and more realistic settings. Another promising direction is investigating models with more than two groups. Even though we chose to focus on applications with two groups in this work, our method can generalize to more than two groups, which is a promising direction for future work.

REFERENCES

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.

Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Chris Burgess and Hyunjik Kim. 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

Tian Qi Chen, Xuechen Li, Roger B. Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 2615–2625, 2018.

Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa A Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. *arXiv preprint arXiv:1906.02589*, 2019.

Emilien Dupont. Learning disentangled joint continuous and discrete representations. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 708–718, 2018.

Partha Ghosh, Mehdi SM Sajjadi, Antonio Vergari, Michael Black, and Bernhard Schölkopf. From variational to deterministic autoencoders. *arXiv preprint arXiv:1903.12436*, 2019.

Irina Higgins, Loïc 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 *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017a.

Irina Higgins, Arka Pal, Andrei A. Rusu, Loïc Matthey, Christopher Burgess, Alexander Pritzel, Matthew Botvinick, Charles Blundell, and Alexander Lerchner. DARLA: improving zero-shot transfer in reinforcement learning. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1480–1490. PMLR, 2017b.

Matthew D Hoffman and Matthew J Johnson. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*, 2016.

Haruo Hosoya. Group-based learning of disentangled representations with generalizability for novel contents. In *IJCAI*, pp. 2506–2513, 2019.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

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

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Jack Klys, Jake Snell, and Richard S. Zemel. Learning latent subspaces in variational autoencoders. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 6445–6455, 2018.

Jean Kossaifi, Linh Tran, Yannis Panagakis, and Maja Pantic. GAGAN: geometry-aware generative adversarial networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 878–887. IEEE Computer Society, 2018.

Zhuo Li, Hongwei Wang, Miao Zhao, Wenjie Li, and Minyi Guo. Deep representation-decoupling neural networks for monaural music mixture separation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Xinhai Liu, Zhizhong Han, Yu-Shen Liu, and Matthias Zwicker. Fine-grained 3d shape classification with hierarchical part-view attentions. *IEEE Transactions on Image Processing*, 2021.

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.

Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. In *Advances in Neural Information Processing Systems*, pp. 14611–14624, 2019a.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pp. 4114–4124. PMLR, 2019b.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. *arXiv preprint arXiv:2002.02886*, 2020.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

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. PMLR, 2014.

Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly supervised disentanglement with guarantees. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Samarth Sinha and Adji B Dieng. Consistency regularization for variational auto-encoders. *arXiv preprint arXiv:2105.14859*, 2021.

Linh Tran, Jean Kossaifi, Yannis Panagakis, and Maja Pantic. Disentangling geometry and appearance with regularised geometry-aware generative adversarial networks. *International Journal of Computer Vision*, 127(6):824–844, 2019.

Frederik Träuble, Elliot Creager, Niki Kilbertus, Anirudh Goyal, Francesco Locatello, Bernhard Schölkopf, and Stefan Bauer. Is independence all you need? on the generalization of representations learned from correlated data. *arXiv e-prints*, pp. arXiv–2006, 2020.

Sjoerd van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. Are disentangled representations helpful for abstract visual reasoning? In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 14222–14235, 2019.

Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5): 1–12, 2019.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1912–1920, 2015.

Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 206–215, 2018.

A GROUPVAE

A.1 JOINT LEARNING OF CONTINUOUS AND DISCRETE GROUPS

The generative model defined in the main Section 4 assumes both content and style representations to be Gaussian distributed. However, many data-generating processes rely on discrete factors which is usually difficult to capture with continuous variables. In these cases, we can define the generative model as

$$p(\boldsymbol{z}_c) = \text{Cat}(\pi), \tag{9}$$
$$p(\boldsymbol{z}_s) = \mathcal{N}(0, \boldsymbol{I}), \tag{10}$$
$$p(\boldsymbol{x}|\boldsymbol{z}_c, \boldsymbol{z}_s) = \text{Bernoulli}(f_\theta(\boldsymbol{z}_c, \boldsymbol{z}_s)). \tag{11}$$

For inference, we use a Gumbel-Softmax reparameterization Jang et al. (2017); Maddison et al. (2017), a continuous distribution on the simplex that can approximate categorical samples for $\boldsymbol{z}_c$. Similar to the KL divergence between two Normal distributions, the KL divergence between two Categorical distributions can also be computed in closed form.

A.2 CLOSED-FORM SOLUTIONS FOR THE KL REGULARIZATION

In the case of both $\boldsymbol{z} \sim \mathcal{N}(\mu, \sigma^2)$ and $\boldsymbol{z}' \sim \mathcal{N}(\mu', \sigma'^2)$ being factorized Gaussian distributions, the KL regularization has the analytical solution

$$\text{KL}(\boldsymbol{z}||\boldsymbol{z}') = \log \frac{\sigma'}{\sigma} + \frac{\sigma^2 + (\mu - \mu')^2}{2\sigma'^2} - \frac{1}{2}. \tag{12}$$

In the case of $\boldsymbol{z}$ Categorical(π) and $\boldsymbol{z}'$ Categorical(π'), the KL has the analytical solution

$$\text{KL}(\boldsymbol{z}||\boldsymbol{z}') = \sum_i \pi_i \log \frac{\pi_i}{\pi'_i}. \tag{13}$$

B ANALYSIS OF EXISTING GROUP-DISENTANGLEMENT APPROACHES

In this Section, we give further details about the content approximate posterior proposed by Bouchacourt et al. Bouchacourt et al. (2018), and Hosoya Hosoya (2019). Further, we analyze the proposed approaches and show its limitations.

B.1 MLVAE AND GVAE

As described in Subsection 3, we restrict to two groups and define corresponding latent variables $\boldsymbol{z}_c$ and $\boldsymbol{z}_s$ given observation $\boldsymbol{x}$[6]. However, both works also apply to any number of groups. For paired observations $(\boldsymbol{x}, \boldsymbol{x}')$ with shared group factor c, the loss objectives for MLVAE (Bouchacourt et al. (2018)) and GVAE (Hosoya (2019)) are

$$\mathcal{L}_{\text{MLVAE}} = \mathcal{L}_{\text{pairedVAE}} - \beta\text{KL}(q_\phi(\tilde{\boldsymbol{z}}_c, \boldsymbol{z}_s|\boldsymbol{x})||p(\boldsymbol{z})) - \beta\text{KL}(q_\phi(\tilde{\boldsymbol{z}}_c, \boldsymbol{z}_{s'}|\boldsymbol{x}')||p(\boldsymbol{z})), \tag{14}$$
$$\mathcal{L}_{\text{GVAE}} = \mathcal{L}_{\text{pairedVAE}} - \beta\text{KL}(q_\phi(\bar{\boldsymbol{z}}_c, \boldsymbol{z}_s|\boldsymbol{x})||p(\boldsymbol{z})) - \beta\text{KL}(q_\phi(\bar{\boldsymbol{z}}_c, \boldsymbol{z}_{s'}|\boldsymbol{x}')||p(\boldsymbol{z})). \tag{15}$$

The loss objectives $\mathcal{L}_{\text{GVAE}}$ and $\mathcal{L}_{\text{MLVAE}}$ are very similar. The only exceptions are the group approximate posteriors, $\bar{\boldsymbol{z}}_c$ for $\mathcal{L}_{\text{GVAE}}$ and $\tilde{\boldsymbol{z}}_c$ for $\mathcal{L}_{\text{GVAE}}$.

Bouchacourt et al. (2018) assume the group approximate posterior to be a product of the individual approximate posteriors sharing the same group $\bar{\boldsymbol{z}}_c$

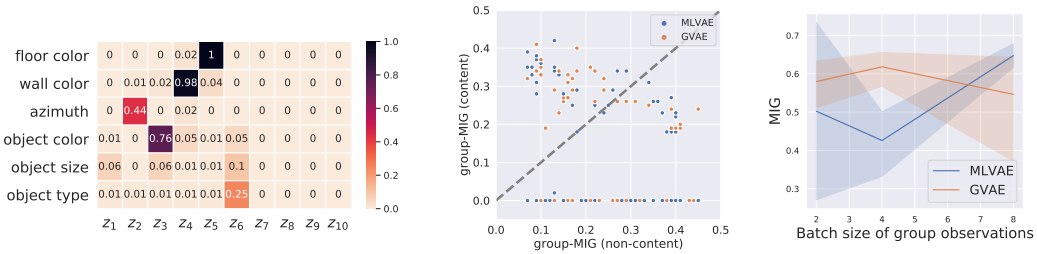
$$\bar{\boldsymbol{z}}_c \sim \mathcal{N}(\mu_{\phi,c}(\boldsymbol{x}), \text{diag}(\sigma^2_{\phi,c}(\boldsymbol{x})) \cdot \mathcal{N}(\mu_{\phi,c}(\boldsymbol{x}'), \text{diag}(\sigma^2_{\phi,c}(\boldsymbol{x}'))). \tag{16}$$

The product of two or more Normal distributions is Normal distributed, and thus the KL term can still be calculated in closed-form.

Hosoya (2019) uses an empirical average over the parameters of the individual approximate posteriors. The group approximate posterior is defined as

$$\tilde{z}_c \sim \mathcal{N}(0.5 \cdot \mu_{\phi,c}(\boldsymbol{x}) + 0.5 \cdot \mu_{\phi,c}(\boldsymbol{x}'), 0.5 \cdot \text{diag}(\sigma^2_{\phi,c}(\boldsymbol{x})) + 0.5 \cdot \text{diag}(\sigma^2_{\phi,c}(\boldsymbol{x}'))). \tag{17}$$

[6] In similar fashion, we define two latent variables $\boldsymbol{z}'_c$ and $\boldsymbol{z}'_s$ for observation $\boldsymbol{x}'$.



(a) 3DShapes: MI between latent dimensions and factors of variation of a trained GVAE model with MIG $= 0.55$ and group-MIG $= 0.44$.

(b) dSprites: group-MIG of content and style information for all hyperparameter runs.

(c) dSprites: MIG w.r.t. different number of shared observations for MLVAE and GVAE.

Figure 5: **Collapse and sensitivity of existing weakly supervised group disentanglement models.** (a) shows mutual information (MI, *higher is better*) for a GVAE model trained on 3DShapes. (b) plots both group-MIG (*higher is better*) w.r.t. content and style information trained on dSprites. (c) plots MIG (*higher is better*) w.r.t. number of shared observations.

B.2 Analysis

Both MLVAE and GVAE enforce disentanglement through the β-regularization in the last two terms of Equations (14, 15). This regularization was also used in β-VAE Higgins et al. (2017a) which regularizes a trade-off between disentanglement and reconstruction. The two KL terms in Equations (14, 15) can be decomposed similar to the ELBO and KL decomposition in Hoffman & Johnson (2016); Chen et al. (2018). We consider the objective in Equation (14) averaged over the empirical distribution $p(n)$. Each training sample denoted by a unique index and treated as random variable n. We simplify $q(\boldsymbol{z}|\boldsymbol{x}_n) = q(\boldsymbol{z}|n)$ and refer to $q(\boldsymbol{z}) = \sum_i^N q(\boldsymbol{z}|n)p(n)$ as the aggregated posterior Hoffman & Johnson (2016). We can decompose the first KL in Equation (14)[7] as

$$
\begin{aligned}
&\mathbb{E}_{p(n)}\Big[\mathrm{KL}\big(q(\bar{\boldsymbol{z}}_c, \boldsymbol{z}_s|n)||p(\boldsymbol{z})\big)\Big] \\
&= \mathrm{KL}\big(q(\bar{\boldsymbol{z}}_c, \boldsymbol{z}_s, n)||q(\bar{\boldsymbol{z}}_c, \boldsymbol{z}_s)p(n)\big) + \sum_j \mathrm{KL}\big(q(\bar{\boldsymbol{z}}_j)||p(\boldsymbol{z}_j)\big) \\
&\quad + \underbrace{\mathrm{KL}\big(q(\bar{\boldsymbol{z}}_c)||\prod_j q(\boldsymbol{z}_{c,j})\big)}_{\text{content total correlation}} + \underbrace{\mathrm{KL}\big(q(\boldsymbol{z}_s||\prod_j q(\boldsymbol{z}_{s,j}))\big)}_{\text{style total correlation}},
\end{aligned}
\tag{18}
$$

where $q(\bar{\boldsymbol{z}}) = q(\bar{\boldsymbol{z}}_c, \boldsymbol{z}_s) = q(\bar{\boldsymbol{z}}_c) \cdot q(\boldsymbol{z}_s)$. We show the full derivation in the next Subsection B.3. Minimizing the averaged KL between the content and style latent variables ($q(\bar{\boldsymbol{z}}_c, \boldsymbol{z}_s|n)$) and the prior $p(\boldsymbol{z})$ also leads to minimization of the total correlation of content variables and style variables (the last two terms in Equation (18)). The total correlation quantifies the amount of information shared between multiple random variables, i.e., low total correlation indicates high independence between the variables. Even though this objective motivates disentangled content and style representations, the group representation depends on the number of samples used for the averaging. Further, both Bouchacourt et al. (2018) and Hosoya (2019) only average over the content group. There are no structural nor optimization constraints that prevent the style latent variable from encoding all factors of variation.

Sensitivity to group batch size. MLVAE and GVAE use different types of averaging over group latent variables. In realistic settings, always having a certain number of observations that share the same group variations can be difficult. For instance, when training MLVAE and GVAE with dSprites, the performance and its variance is correlated with the number of shared observations. We visualized these findings in Figure 5 (c).

[7] We can decompose the KL of GVAE in Equation (15) similarly.

Visualization of collapse. We visualize such behavior in Figure 5 (a) on a GVAE model trained on 3DShapes with two groups of variations $c = \{\text{object color, object size and object type}\}$ and $s = \{\text{floor color, wall color, azimuth}\}$. Ideally, $\boldsymbol{z}_1 - \boldsymbol{z}_5$ contains high mutual information with group factors s and $\boldsymbol{z}_6 - \boldsymbol{z}_{10}$ contains high mutual information with group factors c. However, most information is captured in $\boldsymbol{z}_1 - \boldsymbol{z}_5$, whereas only a little information about object type is contained in $\boldsymbol{z}_6 - \boldsymbol{z}_{10}$.

B.3 KL Decomposition

Here, we show the full derivation for Equation (18). For a given group g the KL decompose as follows:

$$\mathbb{E}_{p(n)}\Big[\mathrm{KL}\big(q(\bar{\boldsymbol{z}}_c, \boldsymbol{z}_s|n)||p(\boldsymbol{z})\big)\Big] \tag{19}$$

$$= \mathbb{E}_{p(n)}\Big[\mathbb{E}_{q(\bar{\boldsymbol{z}}_c,\boldsymbol{z}_s|n)}\big[\log q(\bar{\boldsymbol{z}}_c, \boldsymbol{z}_s|n) - \log p(\boldsymbol{z})\big]\Big] \tag{20}$$

$$= \mathbb{E}_{p(n)}\Big[\mathbb{E}_{q(\bar{\boldsymbol{z}}_c,\boldsymbol{z}_s|n)}\big[\log q(\bar{\boldsymbol{z}}_c, \boldsymbol{z}_s|n) - \log p(\boldsymbol{z}) + \underbrace{\log q(\bar{\boldsymbol{z}}_c, \boldsymbol{z}_s) - \log q(\bar{\boldsymbol{z}}_c, \boldsymbol{z}_s)}_{=0} + \underbrace{\log \prod_i q(\bar{\boldsymbol{z}}_i) - \log[\prod_i q(\bar{\boldsymbol{z}}_{c,i}) \prod_j q(\boldsymbol{z}_{s,j})]}_{=0}\big]\Big] \tag{21}$$

$$= \mathbb{E}_{q(\bar{\boldsymbol{z}}_c,\boldsymbol{z}_s,n)}\left[\log \frac{q(\bar{\boldsymbol{z}}_c, \boldsymbol{z}_s|n)}{q(\bar{\boldsymbol{z}}_c, \boldsymbol{z}_s)}\right] + \mathbb{E}_{q(\bar{\boldsymbol{z}}_c,\boldsymbol{z}_s)}\left[\log \frac{q(\bar{\boldsymbol{z}}_c, \boldsymbol{z}_s)}{\prod_i q(\bar{\boldsymbol{z}}_{c,i}) \prod_j q(\boldsymbol{z}_{s,j})}\right] + \mathbb{E}_{q(\bar{\boldsymbol{z}})}\left[\log \frac{\prod_i q(\bar{\boldsymbol{z}}_i)}{p(\boldsymbol{z})}\right] \tag{22}$$

$$= \mathbb{E}_{q(\bar{\boldsymbol{z}}_c,\boldsymbol{z}_s,n)}\left[\log \frac{q(\bar{\boldsymbol{z}}_c, \boldsymbol{z}_s|n)p(n)}{q(\bar{\boldsymbol{z}}_c, \boldsymbol{z}_s)p(n)}\right] + \mathbb{E}_{q(\bar{\boldsymbol{z}}_c)}\left[\log \frac{q(\bar{\boldsymbol{z}}_c)}{\prod_i q(\bar{\boldsymbol{z}}_{c,i})}\right] + \mathbb{E}_{q(\boldsymbol{z}_s)}\left[\log \frac{q(\bar{\boldsymbol{z}}_s)}{\prod_j q(\boldsymbol{z}_{s,j})}\right] + \mathbb{E}_{q(\bar{\boldsymbol{z}})}\left[\sum_i \log \frac{q(\bar{\boldsymbol{z}}_i)}{p(\boldsymbol{z}_i)}\right] \tag{23}$$

$$= \underbrace{\mathrm{KL}\big(q(\bar{\boldsymbol{z}}_c, \boldsymbol{z}_s, n)||q(\bar{\boldsymbol{z}}_c, \boldsymbol{z}_s)p(n)\big)}_{\text{Index-code MI}} + \underbrace{\mathrm{KL}\big(q(\bar{\boldsymbol{z}}_c)||\prod_i q(\boldsymbol{z}_{c,i})\big)}_{\text{content total correlation}} + \underbrace{\mathrm{KL}\big(q(\boldsymbol{z}_s||\prod_i q(\boldsymbol{z}_{s,i}))\big)}_{\text{style total correlation}} + \underbrace{\sum_i \mathrm{KL}\big(q(\bar{\boldsymbol{z}}_i)||p(\boldsymbol{z}_i)\big)}_{\text{Dimension-wise KL}}, \tag{24}$$

where $p(n)$ denote the empirical data distribution.

C Experimental Setup

C.1 Disentanglement Study

All hyperparameters for optimization and model architectures are listed in Table 4. We compare our approach, GroupVAE, to four different models: β-VAE Higgins et al. (2017a), AdaGVAE Locatello et al. (2020), MLVAE Bouchacourt et al. (2018) and GVAE Hosoya (2019). To fairly compare all models, we used the same architecture and optimization settings for all models and only varied the range of the regularization. We ran five experiments for every hyperparameter set with different random seeds ($= [0, 1, 2, 3, 4]$). In total, we ran 240 experiments. Each experiment ran on GPU clusters consisting of Nvidia V100 or RTX 6000 for approximately 2-3 hours.

Datasets and group sampling. We evaluated our approach on three datasets: 3D Cars Reed et al. (2014), 3D Shapes Burgess & Kim (2018) and dSprites Matthey et al. (2017). All datasets contain

Parameters	Values
Batch size	64
Latent dimension	10
Optimizer	Adam
Adam: beta1	0.9
Adam: beta2	0.999
Adam: epsilon	$1e$-8
Adam: learning rate	$5e$-4
Training iterations	300,000

(a) Common hyperparameters.

Architecture	
$q_\phi(\boldsymbol{z}\|\boldsymbol{x})$	Conv $32 \times 4 \times 4$ (Stride 2), ReLU act., Conv $32 \times 4 \times 4$ (Stride 2), ReLU act., Conv $64 \times 4 \times 4$ (Stride 2), ReLU act., Conv $64 \times 4 \times 4$ (Stride 2), ReLU act., FC 256, ReLU act., FC 2×10
$p_\theta(\boldsymbol{x}\|\boldsymbol{z})$	FC 1024, ReLU act., Reshape (64, 4, 4), TransposeConv $64 \times 4 \times 4$ (Stride 2), ReLU act., TransposeConv $32 \times 4 \times 4$ (Stride 2), ReLU act., TransposeConv $32 \times 4 \times 4$ (Stride 2), ReLU act., TransposeConv $3 \times 4 \times 4$ (Stride 2)

(b) Common model architectures.

Model	Parameter	Values
β-VAE Higgins et al. (2017a)	β	$[1, 2, 4, 6, 8, 16]$
AdaGVAE Locatello et al. (2020)	β	$[1, 2, 4, 6, 8, 16]$
MLVAE Bouchacourt et al. (2018)	β	$[1, 2, 4, 6, 8, 16]$
GVAE Hosoya (2019)	β	$[1, 2, 4, 6, 8, 16]$
GroupVAE	λ	$[1, 2, 8, 16, 32, 64]$

(c) Model hyperparameters.

Table 4: **Experimental setup for the disentanglement study.** We list hyperparameters, model architectures and hyperparamter common to the disentanglement study.

images of size 64×64 with pixels normalized between 0 and 1. For training, given observations $\boldsymbol{x}$ and groups $g_1, \ldots, g_m$, we sample uniformly g from all groups and the observation $\boldsymbol{x}'$ uniform from all observations which share the same group values as $\boldsymbol{x}$.

Evaluating disentanglement. In addition to comparing group disentanglement, we also used MIG Chen et al. (2018) to compare the models' ability to disentangle all factors of variation. Chen et al. (2018) proposed MIG as an unbiased and hyperparameter-dependent evaluation metric to measure the mutual information between each ground truth factor and each dimension in the computed representation. The MIG is calculated as the average difference between the highest and second-highest normalized mutual information of each factor. The score is computed as

$$\text{MIG} = \frac{1}{K} \sum_{1}^{K} \frac{1}{H(v_k)} \big(I_n(\boldsymbol{z}_{j^{(k)}}; v_k) - \max_{j \neq j^{(k)}} I_n(\boldsymbol{z}_j; v_k) \big), \tag{25}$$

where $j^{(k)} = \arg\max_j I_n(\boldsymbol{z}_j; v_k)$ and K is the number of known factors.

C.2 Fairness

We ran five experiments for every hyperparameter set with different random seeds ($= [0, 1, 2, 3, 4]$). In total, we ran 550 experiments. Each experiment ran on GPU clusters consisting of Nvidia V100 or RTX 6000 for approximately 2-3 hours.

Models For the fair classification experiments we used the same common hyperparameters and model architecture as in the disentanglement studies (Table 4 (a) and (b)) for GroupVAE, GVAE and MLVAE. In addition, we implemented two simple baselines, an Multi-Layer Perceptron (MLP) and a Convolutional Neural Network (CNN). The architecture for these two models are described in Table 5. For the supervised fair classification, we implemented FFVAE Creager et al. (2019) with the same encoder and decoder networks as in Table 4 (b) and the FFVAE discriminator as in Table 5. The baselines are trained with a cross-entropy loss between the logits of the network and the binary label "HeavyMakeup". We used different number of latent dimensions which is shown in Table 5 (c).

Architecture	
FFVAE discriminator	FC 1000, LeakyReLU(0.2) act., FC 1000, LeakyReLU(0.2) act., FC 1000, LeakyReLU(0.2) act., FC 1000, LeakyReLU(0.2) act., FC 1000, LeakyReLU(0.2) act., FC 2
f_{MLP}	FC 128, ReLU act., FC 128, ReLU act., FC 128, FC 2
f_{CNN}	Conv $1 \times 6 \times 5$, ReLU act., MaxPool Conv $6 \times 16 \times 5$, ReLU act., MaxPool, FC 120, ReLU act., FC 84, ReLU act., FC 2

(a) Additional model architecture.

Model	Parameter	Values
FFVAE	α	$[0, 1, 100, 300, 500, 1000]$
	γ	$[10, 20, 30, 40, 50, 100]$

(b) Additional model hyperparameter.

Dataset	Parameters	Values
CelebA	Latent dimensions [sensitive, non-sensitive]	$[[3, 37], [40, 40]]$
dSpritesUnfair	Latent dimensions [sensitive, non-sensitive]	$[[5, 5]]$

(c) Dataset-specific hyperparameters.

Table 5: **Experimental settings for fair classification.** We list hyperparameters of FFVAE and the MLP and CNN baselines.

Sensitive and non-sensitive latent variables. Similar to the content and style disentanglement setup, we define two groups, sensitive and non-sensitive. GroupVAE can be optimized to learn from weakly supervised observations sharing either sensitive or non-sensitive group values. FFVAE Creager et al. (2019) can be seen as the supervised approach of learning sensitive and non-sensitive representations. FFVAE maximizes the ELBO objective (reconstruction loss and KL divergence between approximate posterior and prior). In addition, the objective regularizes the discriminative ability of the sensitive latent variable with α in a supervised manner (*how well can the model classify sensitive labels from sensitive latent variable?*) and the disentanglement with γ (*how well is the sensitive latent variable disentangled from the non-sensitive latent variable?*).

Datasets. For comparability with FFVAE Creager et al. (2019), we used similar dataset settings for CelebA Li et al. (2018) and dSpritesUnfair. Both datasets contain images with pixels normalized between 0 and 1. We used the pre-defined train, validation, and test split of CelebA Li et al. (2018), whereas in dSpritesUnfair we use a random split of 80% train, 5% validation, and 15% test.

dSpritesUnfair. dSpritesUnfair is a modified version of dSprites Matthey et al. (2017). The two components are the binarization of the factors of variation and biased sampling. dSprites contains images which are described by five factors of variation. We binarized the factors of variations following these criterion Creager et al. (2019):

- Shape ≥ 1
- Scale ≥ 3
- Rotation ≥ 20
- X-position ≥ 16
- Y-position ≥ 16

Similar to Träuble et al. (2020), we enforce correlations between shape and x-position through a biased sampling. In the training set, we sample these two factors from a joint distribution

$$p(s, x) \propto \exp\Big(- \frac{(s - x)^2}{2\sigma^2} \Big), \tag{26}$$

where σ determines the strength of the.correlation and is set to $\sigma = 0.2$ in our experiments. The smaller σ, the higher the correlation between the two factors.

Model selection. As shown in Creager et al. (2019), there is a trade-off between classification accuracy and demographic parity. Thus, model selection based on only one of these metrics compromises the other. We propose to use the difference between the two metrics as a way to do model

selection. We coin this metric FairGap (FG) and define it as

$$FG = \underbrace{\mathbb{E}[\bar{y} = y]}_{\text{Accuracy}} - \frac{1}{|a|} \sum_{a} \underbrace{|\mathbb{E}[\bar{y} = 1|a = 1] - \mathbb{E}[\bar{y} = 1|a = 0]|}_{\text{demographic parity}}. \tag{27}$$

FG is high if accuracy is high and the average demographic is low, resulting in a fair classifier. We select the model on the test set of CelebA and dSprites based on the FG of the validation set.