Discovering Interpretable Variations from Perceptual Consistency and Orthogonality

Anonymous Author(s)
Affiliation
Address
email

Abstract

We argue that the identification of interpretable variations for visual data depends on perceptual-level vision commonsense. This commonsense enables the ability to sense the characteristics of different variations, and summarize some of them as interpretable based on some biases. We thus leverage a model with such vision commonsense to guide the discovery of interpretable variations. This commonsense model which can perceive variations is named as variation perceiver, and we instantiate it with a basic pretrained classifier in this paper. We then impose two inductive biases, namely perceptual consistency and orthogonality, as objectives to encourage the discovery process based on the variation perceiver. The two inductive biases state that the variations belonging to the same type should be perceptually similar, and those belonging to different types should be perceptually orthogonal. We instantiate our method with pretrained GANs and show that the discovered directions can be used to not only manipulate images, but learn state-of-the-art disentangled representations. The effectiveness of our proposed method is validated on various datasets quantitatively and qualitatively.

1 Introduction

Deep generative models such as GANs Goodfellow et al. [2014] and VAEs Kingma and Welling [2013] are becoming more and more advanced in generating high-resolution and photo-realistic images Karras et al. [2018, 2020a], Brock et al. [2019], Vahdat and Kautz [2020]. One important line of work that augments these generative models is to assign them with interpretability Yan et al. [2016], Chen et al. [2016], Higgins et al. [2017], Bau et al. [2019], Shen et al. [2019], which lays the foundations for downstream applications such as controllable image generation Kulkarni et al. [2015], Lample et al. [2017], Lee et al. [2018], Xing et al. [2019], domain adaptation Peng et al. [2019], Cao et al. [2018], machine learning fairness Creager et al. [2019], Locatello et al. [2019], and etc. This task can usually be achieved with supervised learning on datasets with labels of attributes Kingma et al. [2014], Dosovitskiy et al. [2014], Kulkarni et al. [2015], Lample et al. [2017], Shen et al. [2019]. Another branch of research is to encourage models to discover these interpretable attributes without supervision Chen et al. [2016], Higgins et al. [2017], Shen and Zhou [2021]. This paper focuses on the later task.

A conventional way to achieve this unsupervised learning goal is by learning a disentangled generative model so that the dimension-wise change in the latent space corresponds to a single type of variation in the data space Bengio et al. [2012]. This task can be tackled by imposing statistical independence assumption Higgins et al. [2017], Burgess et al. [2018], Kim and Mnih [2018], Chen et al. [2018], or maximizing informativeness of the latents Chen et al. [2016], Jeon et al. [2018], Lin et al. [2020], Zhu et al. [2021]. The models learned from this perspective usually suffer from downgraded generation quality. Recently, another fashion of learning interpretable variations is by discovering semantic directions from a pretrained generator (usually a pretrained GAN) Voynov and Babenko [2020], Härkönen et al. [2020], Shen and Zhou [2021]. Via this modeling, the extraction of semantic factors is totally separated from the generation task, leaving the generation quality of the generative model untouched. In this paper, we also show that conventional disentangled models can be derived from the semantic discovery models.

Existing semantic discovery models have some limitations. Methods from Jahanian et al. [2020] and Plumerault et al. [2020] can only discover variations of simple transformations such as scaling and translation. While methods in Härkönen et al. [2020], Shen and Zhou [2021] can discover some general semantics, they rely on manual selection of the layers in the generator to which their methods can be applied. In Voynov and Babenko [2020], a classifier is used to encourage the discovered semantics to be distinguishable. However the classifier can be trained to grouping non-interpretable variations, leading to sub-optimal effectiveness. Additionally, none of these models leverage the spatial information in images to enforce the discovery of localized interpretable variations.

We rethink this semantic discovery problem from the perspective of how we humans define interpretable variations. Firstly, it should be admitted that interpretability is a concept that essentially requires perceptual-level evaluation, i.e. we cannot tell if something is interpretable before we understand it semantically. Secondly, when trying to define interpretable factors on a new domain, we usually choose the variations that can be easily recognized and distinguished by commonsense. It matches the intuition behind interpretability as being *widely-understandable*. The term *commonsense* refers to some knowledge that is general and usually gained from other domains (experience). We propose to simulate this procedure with deep learning models by using a network to evaluate the variations based on vision commonsense, and use it to guide the discovery of semantic variations in the new domain. Note that the models with vision commonsense can usually refer to networks that are pretrained on large scale data with or without supervision Krizhevsky et al. [2017], Simonyan and Zisserman [2015], Chen et al. [2020a,b].

In this paper, we construct a model called variation perceiver with a pretrained classifier, and use the extracted general perceptual features to support interpretable variation discovery. As interpretable variations are supposed to be easily recognizable and distinguishable, we propose two constraints, namely perceptual consistency and orthogonality, as loss functions to encourage the learning of our discovery model. The perceptual consistency constraint enforces the variations of the same type to be perceptually parallel. The orthogonality constraint ensures that different variations are perceptually orthogonal. With the interpretable directions discovered on a pretrained GAN, we show that conventional disentangled representations can be derived by (1) generating a dataset of factor-changed image-pairs, and (2) training a disentanglement model on this dataset. This also enables the quantitative evaluation of the semantic discovery models. We demonstrate that our method can achieve successful image editing on various datasets, and learn state-of-the-art disentanglement models.

2 Related Work

Generative Adversarial Network. GAN Goodfellow et al. [2014] is a type of generative model which comprises a generator and a discriminator trained under an adversarial strategy, and state-of-the-art GANs Karras et al. [2018], Brock et al. [2019], Karras et al. [2020b,a] have been developed to synthesize photo-realistic images in high resolution. Besides their advanced capability in generation, it has been shown that well-trained GANs have interpretable internal representations Bau et al. [2019].

Semantics in GANs. Conventionally we can enable GANs with semantic controls by providing labels during training Mirza and Osindero [2014], Odena et al. [2017], Lample et al. [2017]. Recently, it has been shown that semantic directions can be discovered in a post-hoc fashion using a pretrained GAN Shen et al. [2019]. This can be achieved with labels Shen et al. [2019], or with predefined variations Plumerault et al. [2020], Jahanian et al. [2020], or even without supervision Voynov and Babenko [2020], Härkönen et al. [2020], Shen and Zhou [2021]. In Voynov and Babenko [2020] a classifier is adopted as a regularization to recognize the manipulated direction in the GAN latent space. In Härkönen et al. [2020], principal components found with a sampling strategy in the latent feature space is assumed to be interpretable directions. In Shen and Zhou [2021], the eigenvectors of the projection matrices in generators are computed as directions of semantics. Unlike existing methods, we tackle the semantic direction discovery problem by considering essential properties of interpretability.

Unsupervised Disentanglement Learning. Another popular branch of work to discover interpretable variations in generative models is by learning disentangled representations Higgins et al. [2017], Chen et al. [2016]. In this setting, various regularization methods have been proposed based on different assumptions, e.g. statistical independence Burgess et al. [2018], Kumar et al. [2018], Kim and Mnih [2018], Chen et al. [2018], informativeness Chen et al. [2016], Jeon et al. [2018], Zhu et al. [2021], and separability Lin et al. [2020], Zhu et al. [2020]. These models usually come at the cost of downgraded generation quality compared to the backbone generative model. Unlike the semantic direction discovery setting, a disentangled representation provides semantic embeddings of data samples rather than just semantic directions, which enables quantitative evaluation if the ground-truth factors are available Kim and Mnih [2018], Kumar et al. [2018], Chen et al. [2018], Eastwood and Williams [2018]. In this work, we show the that disentangled representations can be derived based on the discovered semantic directions.

3 Method

In this section, we first introduce our proposed method in 3.1, including the variation perceiver and the proposed constraints for discovering interpretable variations. Then we introduce the instantiation of our semantic direction discovery method in 3.2. Finally a simple method to train a disentangled model is introduced in 3.3.

3.1 Discovering Interpretable Variations

As mentioned in the introduction, we propose to model the discovery of interpretable variations by simulating how humans determine if some variations are interpretable. We summarize there are two requirements to achieve this. The first one is a vision commonsense system that can perceive variations at a semantic level. The extracted semantic encodings of the variations should ideally reflect their relation in the abstract commonsense space, in which lies humans' sensation about these variations. It can be assumed that the better the vision commonsense system matches the humans' vision commonsense system, the more accurate the semantic encoding would be. The second requirement is the criteria of interpretability, where we assume the target variations to be easily recognizable and distinguishable. The first property indicates that each target variation should have consistent feature patterns, while the second property indicates any feature patterns of different variations should be orthogonal. The modeling of the first requirement is introduced in Sec. 3.1.1 and the second one is in Sec. 3.1.2.

3.1.1 Variation Perceiver.

We construct a variation perceiver to obtain the semantic representation of variations. For vision tasks, variations are naturally described by sequences of images. In this paper, we consider the simplest scenario where two images are used. To extract representations of variations, we need a model which maps pairs of images to a vector space $E : \mathcal{X} \times \mathcal{X} \to \mathcal{V}$, where $\mathcal{X}$ denotes the image space and $\mathcal{V}$ denotes the variation representation space. Ideally, the perceiver can be pretrained on a task targeting at interpretability learning, e.g. doing predictions on whether a variation is interpretable. This is a special case of commonsense pretraining where the interpretability is directly used as the pretext task. However, this is infeasible in practice because it requires the labeling of individual variations (image pairs) by whether they are interpretable which is concentration-demanding.

Instead, we consider using generally pretrained networks to extract features that are analogous to vision commonsense of humans. This meets the requirement of perceptual-level understanding of the images, while avoiding the injection of any special knowledge about the new domain. In this paper, we use a basic CNN classifier (e.g. AlexNet Krizhevsky et al. [2017]) to extract feature maps of the paired images, and use their difference to represent the variation $E(\boldsymbol{x_1}, \boldsymbol{x_2}) = C(\boldsymbol{x_1}) - C(\boldsymbol{x_2}), \boldsymbol{x_1}, \boldsymbol{x_2} \in \mathcal{X}$, where C denotes the CNN feature extractor. The underlying assumption is that the perceiver should produce an encoding of the variations which analogously reflects the sensation in humans' perceptual system. We define the variation representation on feature maps, where the spatial information enables the discovery of more localized and fine-grained variations. It is also possible to adopt video-based feature extractors Simonyan and Zisserman [2014], Tran et al. [2015], Carreira and Zisserman

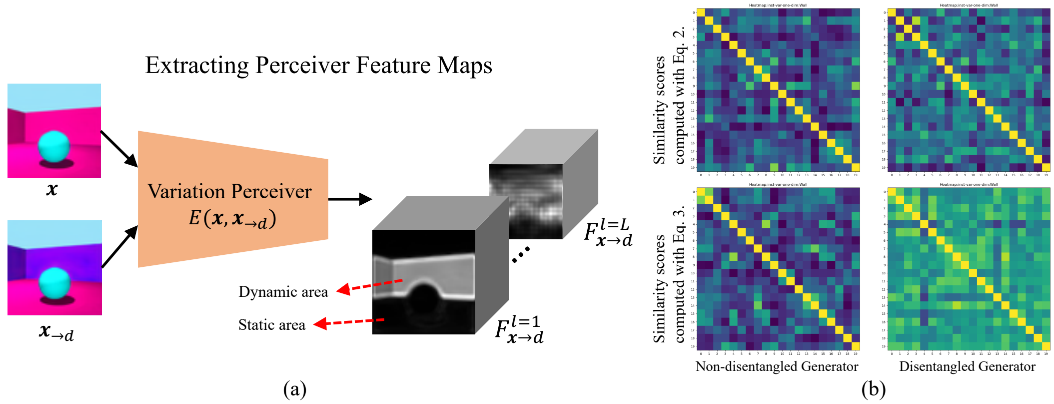

Figure 1: (a). Extracting variation feature maps with a variation perceiver. The extracted feature maps are not evenly activated. For dynamic areas, the norm of each pixel is large while in static areas it is small. (b) Comparing the pair-wise similarity scores computed on 20 random variation samples (of wall color changes as in (a)) using Eq. 2 and 3 on a disentangled and a non-disentangled generator respectively. The constraint defined by Eq. 3 can more effectively distinguish the disentangled and non-disentangled models.

[2017], Alayrac et al. [2020] to construct the variation perceiver as they possess the commonsense of temporal coherency, which is beneficial to the encouragement of variation smoothness. We leave this improvement direction for future work.

3.1.2 Assumptions for Interpretable Variations.

After we obtain representations of variations, we can impose the assumptions of interpretability as constraints to encourage the discovery of interpretable variations. In this paper, we assume the interpretable variations satisfy the consistency and orthogonality constraints on the extracted feature maps. Note that these assumptions are heuristic and may not work on some datasets of special domains (e.g. medical images), but as current semantic discovery models are usually applied to natural vision domains, these assumptions usually hold for discovering common interpretable variations.

Consistency. We expect the variations belonging to a same type to be similar to each other (so that they are easily recognizable). Assuming we have a generator mapping a latent code to an image $G : \mathcal{Z} \to \mathcal{X}$, we need to find a set of directions $\{\boldsymbol{v}_{d=1..m}\}$ in the latent space $\mathcal{Z}$ that control interpretable variations in the image space. Based on the variation perceiver defined earlier, we can obtain a variation representation for a step of change in a direction $\boldsymbol{v}_d$:

$$[F^l_{\boldsymbol{x}\to d}]_{l=1..L} = E(\boldsymbol{x}, \boldsymbol{x}_{\to d}) = E(G(\boldsymbol{z}), G(\boldsymbol{z} + \alpha \boldsymbol{v}_d)), \quad 1 \le d \le m, \tag{1}$$

where we use $\boldsymbol{x}_{\to d} = G(\boldsymbol{z} + \alpha \boldsymbol{v}_d))$ to briefly denote the changed image caused by the d^{th} direction move. The coefficient α denotes the step size in the latent space, and we keep it a small constant. Note that there are L layers of feature maps extract by the variation perceiver, and the values of the variation representations depend on not only the chosen direction d but the current image sample $\boldsymbol{x}$. To enforce consistency, we force the variation representations at different data samples caused by the same moving direction to be similar:

$$\{\boldsymbol{v}\}^*_{\text{cons}} = \arg\max_{\{\boldsymbol{v}\}} \texttt{Cons}_{\texttt{avg}}(\{\boldsymbol{v}\}) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y},l,d}\big[\frac{1}{S}\sum_{s=1}^{S} \texttt{Sim}^2(F^{ls}_{\boldsymbol{x}\to d}, F^{ls}_{\boldsymbol{y}\to d})\big], \tag{2}$$

where s indexes the spatial positions of the feature maps, and $\texttt{Sim}$ is the cosine similarity function. $\boldsymbol{x}$ and $\boldsymbol{y}$ are two different image samples. The Eq. 2 computes an aggregated similarity measurement between feature maps by averaging across spatial positions. However, in practice the activated variation features are not evenly distributed across spatial positions, i.e. they are usually more significant in certain areas than others. In Fig. 1 (a) we illustrate this phenomenon with two generated images (generator trained on 3DShapes dataset Kim and Mnih [2018]). The two images differ only in wall color while keeping other factors unchanged. This leads to the uneven activation in the perceiver feature maps, where the activations in static areas are much lower than dynamic areas. The even-aggregation strategy defined in Eq. 2 is not suitable to this case since activations from static areas should not contribute to the measurement of variation consistency. Based on this concern, we use a natural mask derived from the L2-norm of the activations to realize a weighted aggregation of the similarity scores at different spatial locations:

$$\{\boldsymbol{v}\}^*_{\text{cons}} = \arg\max_{\{\boldsymbol{v}\}} \texttt{Cons}_{\texttt{mask}}(\{\boldsymbol{v}\}) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y},l,d}\big[\frac{1}{\bar{q}(s)}\sum_{s=1}^{S} q(s)\texttt{Sim}^2(F^{ls}_{\boldsymbol{x}\to d}, F^{ls}_{\boldsymbol{y}\to d})\big], \tag{3}$$

$$\bar{q}(s) = \sum_{s=1}^{S} q(s), \quad q(s) = \texttt{norm}(F^{s}_{\boldsymbol{x}\to d}) \times \texttt{norm}_{\texttt{max}}(F^{s}_{\boldsymbol{x}\to d}), \tag{4}$$

where $\texttt{norm}(F^{s}_{\boldsymbol{x}\to d})$ computes the L2-norm of the perceiver feature in $F_{\boldsymbol{x}\to d}$ at position s. This weighted aggregation forces the measurement to focus more on the dynamic areas in both the two input variation samples, while paying less attention to areas where both variation samples are static. Empirically this version of consistency measurement also leads to more effective results. In Fig. 1 (b), we show a comparison between the pair-wise similarity scores computed by 20 random variation samples (of wall color variation similar to Fig. 1 (a)), using Eq. 2 and Eq. 3 on a non-disentangled generator and a disentangled generator respectively. We see the Eq. 3 can more effectively differentiate disentangled and non-disentangled models, which works as a stronger guidance for encouraging semantic discovery.

Orthogonality. Similarly we define the orthogonality constraint, which encourages the different variations to be perceptually orthogonal evaluated with the variation perceiver:

$$\{\boldsymbol{v}\}^*_{\text{orth}} = \arg\min_{\{\boldsymbol{v}\}} \texttt{Orth}_{\texttt{mask}}(\{\boldsymbol{v}\}) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y},l,d,d'}\big[\frac{1}{\bar{q}(s)}\sum_{s=1}^{S} q(s)\texttt{Sim}^2(F^{ls}_{\boldsymbol{x}\to d}, F^{ls}_{\boldsymbol{y}\to d'})\big], \tag{5}$$

where d, d' represent different directions in the latent space. This contributes to the assumption that different interpretable variations should be easily distinguishable. The overall constraint is to combine the consistency and orthogonality:

$$\{\boldsymbol{v}\}^* = \arg\min_{\{\boldsymbol{v}\}} -\texttt{Cons}_{\texttt{mask}}(\{\boldsymbol{v}\}) + \texttt{Orth}_{\texttt{mask}}(\{\boldsymbol{v}\}). \tag{6}$$

3.2 Model Instantiation

We instantiate this idea with StyleGAN2 Karras et al. [2020a] by discovering interpretable directions in its $\mathcal{W}$ space. For a latent sample $\boldsymbol{w} \in \mathcal{W}$, we use a navigator network to predict the interpretable directions:

$$\{\boldsymbol{v}_{d=1..m}\}_{\boldsymbol{w}} = M(\boldsymbol{w}), \tag{7}$$

where M is a three layer MLP. Using $\mathcal{W}$ also makes it easier to edit real images since it has been shown that real images can be more easily inverted into the $\mathcal{W}$ space than the input $\mathcal{Z}$ space Karras et al. [2020a], Abdal et al. [2019, 2020]. We can then edit a real image by altering the corresponding latent vector in the direction of certain discovered variation:

$$\boldsymbol{x}_{\to d} = G(\boldsymbol{w}^* + \alpha M(\boldsymbol{w}^*)), \quad \boldsymbol{w}^* = G^{-1}(\boldsymbol{x}_{\text{real}}), \tag{8}$$

where G^{-1} denotes the image inversion (projection) process, and $\boldsymbol{w}^*$ is the inverted latent code of the real image $\boldsymbol{x}_{\text{real}}$.

3.3 Disentanglement Measurement

Quantitative evaluation on semantic discovery models is not straightforward. Existing methods include using pretrained attribute-classifiers to measure the attribute altering quality or user-study Shen and Zhou [2021] which are not scalable and sub-reliable. Here we introduce a simple method to leverage the disentanglement metrics by learning a disentangled representation with a semantic discovery model.

To use the off-the-shelf disentanglement metrics Higgins et al. [2017], Kim and Mnih [2018], Eastwood and Williams [2018], Chen et al. [2018], we need an encoder model which maps an image into a disentangled representation space. Unfortunately the semantic discovery models only know the directions to alter attributes but do not know the current attribute-embedding. It is thus required to train such an encoder based on the direction information. We point out that this training task exactly matches the weakly-supervised disentanglement learning setting where the impact on the observation space caused by a subset of generative factors is known Hosoya [2019], Bouchacourt et al. [2018], Locatello et al. [2020], Painter et al. [2020]. We can thus adopt the existing methods on this weakly-supervised setting to train a disentangled encoder for quantitative evaluation. In this paper, we use a simple method named GVAE Hosoya [2019] to achieve this goal.

For implementation, we first generate a dataset of image-pairs with each sample showing a step of change in a discovered direction:

$$\{(\boldsymbol{x}, \boldsymbol{x}_{\rightarrow}.)\} = \{(G(\boldsymbol{z}), G(\boldsymbol{z} + \alpha M(\boldsymbol{z})_d)) \,|\, \boldsymbol{z} \sim p(\boldsymbol{z}), d \sim U_{\text{int}}(1, m)\}, \tag{9}$$

where $p(\boldsymbol{z})$ is the prior distribution of the latent code $\boldsymbol{z}$, and $U_{\text{int}}(1, m)$ is a uniform distribution of m integers to index the latent moving direction. Then we train a GVAE on the generated image-pair dataset (see Appendix 6 for an introduction) and report the disentanglement scores. These scores implicitly reflect the performance of the semantic direction discovery method, i.e. if the discovered directions are disentangled enough, the generated dataset will be easy enough to train a weakly-supervised disentanglement model because the image-pairs will represent clean factor changes. Note that this method also shows a new pipeline to solve the unsupervised disentanglement learning problem, in which we (1) train a generator; (2) discover interpretable directions; (4) generate an image-pair dataset with the discovered directions; and finally (3) train a weakly-supervised disentanglement model.

4 Experiments

We present the semantic discovery results on StyleGAN2 Karras et al. [2020a] pretrained on FFHQ Karras et al. [2020b], CelebA Karras et al. [2018], AFHQ Choi et al. [2020], and LSUN Car, Church, Cat Yu et al. [2015] in Sec. 4.1, and disentangled representation learning results on 3DShapes Kim and Mnih [2018] and DSprites Matthey et al. [2017] in Sec. 4.2.

4.1 Semantic Discovery

Discovered Semantics. In Fig. 2, we show some discovered semantics with pretrained StyleGAN2 models on various datasets.

Effectiveness of Orthogonality Constraint. If we remove the orthogonality constraint, the discovery model will learn to represent a same semantics with multiple directions, since there is no encouragement to force the directions to represent different semantics.

Effectiveness of Consistency Constraint. If we remove the consistency constraint, the discovery model usually learns directions to capture simple variations of overall image changes, leading to higher chance to be orthogonal to each other.

Effectiveness of L2-Mask. Without the L2-mask, the number of semantics discovered is reduced, with many directions converge to capture subtle image changes.

Different Scales of Step Size α. The step size α influences the discovered semantics. If α is too small, the discovered semantics are usually in small scale such as eye-open or smile. If α is too large, the discovered variations are usually of large scale, but may also be entangled. Therefore it should be tuned to a balanced value. It is an improvement direction to adaptively adjust α based on the dataset.

Real Image Editing. Real image editing can be realized by projecting an image into the latent representation space, and use the proposed discovery model to apply semantic direction on it.

Summary. Orthogonality and consistency are compulsory components. The L2-mask enables more localized variations to be discovered. The scale of step size α influence the discovered semantics. When α is small, significant variations are hard to be extracted. This may be solved by providing the landscape information of the $\mathcal{W}$ space, e.g. integrating with eigenvectors.

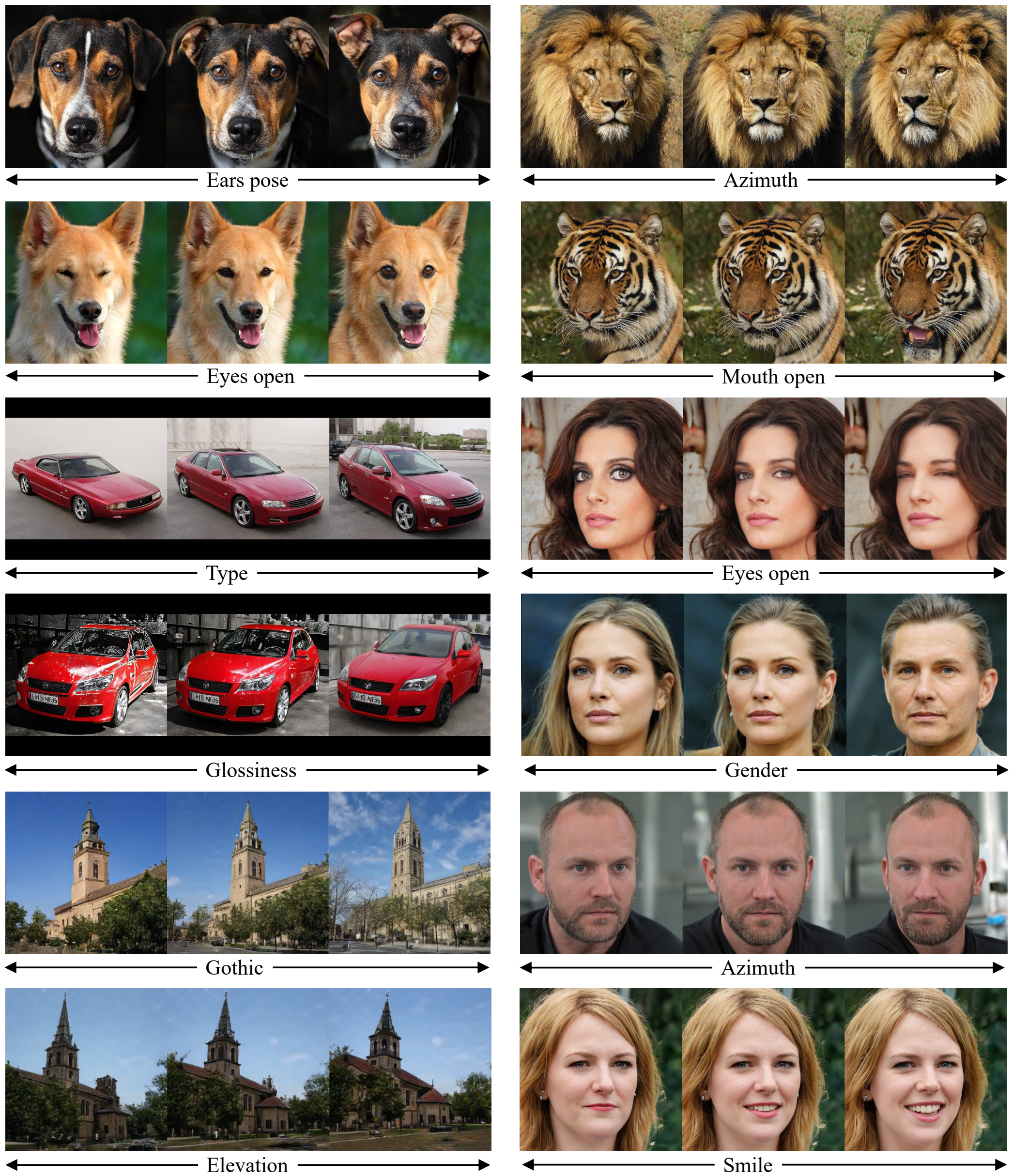

Figure 2: Some of the discovered semantics in StyleGAN2 models pretrained on CelebA-HQ Karras et al. [2018], AFHQ Choi et al. [2020], FFHQ Karras et al. [2020b], and LSUN Car, Church Yu et al. [2015] datasets.

4.2 Disentangled Representation Learning

We conduct this experiment on conventional disentanglement learning datasets 3DShapes Kim and Mnih [2018] and DSprites Matthey et al. [2017]. We consider using existing metrics to quantitatively show the disentanglement property of the discovered variations. The metrics include FactorVAE metric (FVM) Kim and Mnih [2018], Mutual Information Gap (MIG) Chen et al. [2018], DCI metrics Eastwood and Williams [2018], Modularity (MOD) metric Ridgeway and Mozer [2018], and SAP metric Kumar et al. [2018].

Quantitative Ablation Study. State-of-the-art Comparison.

Model	FVM	MIG	DCI-comp	DCI-dis	DCI-info	MOD	SAP
β-VAE	$79.0_{\pm 11.7}$	$35.9_{\pm 18.6}$	$53.3_{\pm 13.4}$	$54.1_{\pm 12.3}$	$78.3_{\pm 5.8}$	$79.3_{\pm 5.9}$	$29.4_{\pm 10.0}$
FacVAE	$79.5_{\pm 7.4}$	$34.6_{\pm 16.4}$	$62.1_{\pm 9.0}$	$66.5_{\pm 8.0}$	$91.8_{\pm 2.8}$	$86.8_{\pm 2.7}$	$33.4_{\pm 9.2}$
Sefa-all	$88.5_{\pm 9.9}$	$22.8_{\pm 8.8}$	$31.9_{\pm 8.8}$	$37.2_{\pm 10.5}$	$71.5_{\pm 7.6}$	$89.7_{\pm 3.2}$	$32.4_{\pm 11.8}$
Ours-Full	$\mathbf{93.8}_{\pm 6.0}$	$\mathbf{45.2}_{\pm 8.0}$	$\mathbf{72.5}_{\pm 7.2}$	$\mathbf{80.1}_{\pm 8.1}$	$\mathbf{95.9}_{\pm 3.5}$	$\mathbf{92.2}_{\pm 4.1}$	$\mathbf{37.0}_{\pm 15.7}$

Table 1: Ablation study of group size on 3DShapes.

5 Conclusion

Inspired by the intuition behind interpretability (being easily recognizable and distinguishable by commonsense), we proposed to discover interpretable variations with deep learning models in an analogous way. We first constructed a variation perceiver with a pretrained network, which simulated the vision commonsense system, to provide a perceptual evaluation on image variations. Then we defined two criteria, namely perceptual consistency and orthogonality, to filter out non-interpretable variations. These two criteria were in practice implemented based on feature maps with cosine similarity measurement, and were used as loss functions to encourage the discovery of semantic directions in the GAN latent space. We pointed out that when the interpretable directions were discovered, disentangled representations can be subsequently learned based on a constructed image-pair dataset in a weakly-supervised learning setting. Empirically we showed the effectiveness of the proposed semantic discovery model, and its state-of-the-art performance on disentangled representation learning.

References

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. In *NIPS*, 2014.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2013.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. 2018.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *CVPR*, 2020a.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. 2019.

Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. In *Neural Information Processing Systems (NeurIPS)*, 2020.

Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. In *ECCV*, 2016.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, 2016.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. *ArXiv*, abs/1907.10786, 2019.

Tejas D. Kulkarni, William F. Whitney, Pushmeet Kohli, and Joshua B. Tenenbaum. Deep convolutional inverse graphics network. In *NIPS*, 2015.

Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. Fader networks: Manipulating images by sliding attributes. In *NIPS*, volume abs/1706.00409, 2017.

Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. *ECCV*, abs/1808.00948, 2018.

Xianglei Xing, Tian Han, Ruiqi Gao, Song-Chun Zhu, and Ying Nian Wu. Unsupervised disentangling of appearance and geometry by deformable generator network. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10346–10355, 2019.

Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko. Domain agnostic learning with disentangled representations. In *ICML*, 2019.

Jinming Cao, Oren Katzir, Peng Jiang, Dani Lischinski, Daniel Cohen-Or, Changhe Tu, and Yangyan Li. Dida: Disentangled synthesis for domain adaptation. *ArXiv*, abs/1805.08019, 2018.

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

Francesco Locatello, Gabriele Abbati, Tom Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. In *NeurIPS*, 2019.

Diederik P. Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *NIPS*, 2014.

Alexey Dosovitskiy, Jost Tobias Springenberg, and Thomas Brox. Learning to generate chairs with convolutional neural networks. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1538–1546, 2014.

Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *CVPR*, 2021.

Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:1798–1828, 2012.

Christopher P. Burgess, Irina Higgins, Arka Pal, Loïc Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in beta-vae. *ArXiv*, abs/1804.03599, 2018.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *ICML*, 2018.

Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, 2018.

Insu Jeon, Wonkwang Lee, and Gunhee Kim. Ib-gan: Disentangled representation learning with information bottleneck gan. *ArXiv*, 2018.

Zinan Lin, Kiran Koshy Thekumparampil, Giulia Fanti, and Sewoong Oh. Infogan-cr: Disentangling generative adversarial networks with contrastive regularizers. *ICML*, 2020.

Xinqi Zhu, Chang Xu, and Dacheng Tao. Where and what? examining interpretable disentangled representations. In *CVPR*, 2021.

Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *International Conference on Machine Learning*, pages 9786–9796. PMLR, 2020.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *Proc. NeurIPS*, 2020.

Ali Jahanian, Lucy Chai, and Phillip Isola. On the "steerability" of generative adversarial networks. In *International Conference on Learning Representations*, 2020.

Antoine Plumerault, Hervé Le Borgne, and Céline Hudelot. Controlling generative models with continuous factors of variations. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=H1laeJrKDB`.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, May 2017. ISSN 0001-0782. doi: 10.1145/3065386. URL `https://doi.org/10.1145/3065386`.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020a.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22243–22255. Curran Associates, Inc., 2020b. URL `https://proceedings.neurips.cc/paper/2020/file/fcbc95ccdd551da181207c0c1400c655-Paper.pdf`.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 2020b.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014. URL `http://arxiv.org/abs/1411.1784`.

Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2642–2651. PMLR, 06–11 Aug 2017. URL `http://proceedings.mlr.press/v70/odena17a.html`.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *ICLR*, 2018.

Xinqi Zhu, Chang Xu, and Dacheng Tao. Learning disentangled representations with latent variation predictability. In *ECCV*, 2020.

Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *ICLR*, 2018.

Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 4489–4497, Washington, DC, USA, 2015. IEEE Computer Society. ISBN 978-1-4673-8391-2. doi: 10.1109/ICCV.2015.510. URL `http://dx.doi.org/10.1109/ICCV.2015.510`.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017.

Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-Supervised MultiModal Versatile Networks. In *NeurIPS*, 2020.

Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4431–4440, 2019. doi: 10.1109/ICCV.2019.00453.

Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8293–8302, 2020. doi: 10.1109/CVPR42600.2020.00832.

Haruo Hosoya. Group-based learning of disentangled representations with generalizability for novel contents. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 2506–2513. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/348. URL `https://doi.org/10.24963/ijcai.2019/348`.

Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 2095–2102. AAAI Press, 2018. URL `https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16521`.

Francesco Locatello, Ben Poole, Gunnar Ratsch, Bernhard Scholkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. *ArXiv*, abs/2002.02886, 2020.

Matthew Painter, Jonathon Hare, and Adam Prugel-Bennett. Linear disentangled representations and unsupervised action estimation, 2020.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

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

Karl Ridgeway and Michael C. Mozer. Learning deep disentangled embeddings with the f-statistic loss. In *NeurIPS*, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See line 120.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] Code will be included when the paper is published.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? **[TODO]**
   (b) Did you mention the license of the assets? **[TODO]**
   (c) Did you include any new assets either in the supplemental material or as a URL? **[TODO]**
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? **[TODO]**
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **[TODO]**
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

6 Introduction of GVAE

abd