OOGAN: Disentangling GAN with One-Hot Sampling and Orthogonal Regularization

Bingchen Liu, Yizhe Zhu, Zuohui Fu, Gerard de Melo, Ahmed Elgammal
Department of Computer Science
Rutgers University
{bingchen.liu, yizhe.zhu, zuohui.fu, gerard.demelo}@rutgers.edu, elgammal@cs.rutgers.edu

Abstract

Exploring the potential of GANs for unsupervised disentanglement learning, this paper proposes a novel GAN-based disentanglement framework with One-Hot Sampling and Orthogonal Regularization (OOGAN). While previous works mostly attempt to tackle disentanglement learning through VAE and seek to implicitly minimize the Total Correlation (TC) objective with various sorts of approximation methods, we show that GANs have a natural advantage in disentangling with an alternating latent variable (noise) sampling method that is straightforward and robust. Furthermore, we provide a brand-new perspective on designing the structure of the generator and discriminator, demonstrating that a minor structural change and an orthogonal regularization on model weights entails an improved disentanglement. Instead of experimenting on simple toy datasets, we conduct experiments on higher-resolution images and show that OOGAN greatly pushes the boundary of unsupervised disentanglement.

1 Introduction

A disentangled representation is one that separates the underlying factors of variation such that each dimension exclusively encodes one semantic feature (Bengio, Courville, and Vincent 2013; Kim and Mnih 2018). While the benefits of the learned representation for downstream tasks is questioned by Locatello et al. (2019), disentangling a Deep Neural Network (DNN) is still of great value in terms of human-controllable data generation, data manipulation and post-processing, and increasing the model interpretability. Moreover, disentanglement learning in an unsupervised manner can effectively highlight the biased generative factors from a given dataset, and yield appealing data-analytic properties. In this work, we focus on unsupervised disentanglement learning using GANs (Goodfellow et al. 2014) on images, which brings substantial advances in tasks such as semantic image understanding and generation, and potentially aids research on zero-shot learning and reinforcement learning (Bengio, Courville, and Vincent 2013; Higgins et al. 2017; Lample et al. 2017; Elgammal et al. 2018; Elhoseiny et al. 2017; Zhu et al. 2018; 2019a; 2019b; Kim et al. 2018).

Recent popular methods to tackle the unsupervised disentangling problem are based on GANs (Goodfellow et al. 2014) or VAEs (Kingma and Welling 2014), and many instantiations of these (Saxe et al. 2018; Gabrié et al. 2018; Alemi et al. 2018; Louizos, Ullrich, and Welling 2017) draw on information-theoretical concepts (Shannon 1948). InfoGAN (Chen et al. 2016) seeks to maximize a **Mutual Information** (MI) lower bound between a sampled conditional vector and the generated data, with the expectation that the generator and discriminator will disentangle the vector with respect to the true underlying factors. InfoGAN-CR (Lin et al. 2019) introduces a contrastive regularizer focusing on forming more desirable latent traversals. In contrast, VAE-based approaches (Esmaeili et al. 2019) attempt to optimize a **Total Correlation** (TC) (Watanabe 1960) objective imposed on the inferred latent vector, which achieves disentanglement by encouraging inter-dimensional independence in the latent vector.

TC-based VAE models have proven fruitful in disentangling. However, there is usually a trade-off between the degree of achievable disentanglement and the data-generating ability of VAE (Kim and Mnih 2018). In practice, VAE struggles significantly when trained on higher-resolution images due to its restricted generative power. Furthermore, it only approximates the TC. Since both the marginal distribution of the learned latent representation and the product of its marginals are intractable in VAE, optimization process is usually implicit and complicated. In contrast, with rapid advances (Zhang et al. 2019; Miyato et al. 2018; Karras, Laine, and Aila 2019), GANs have become more stable to train, and their generative power has become unparalleled on high-resolution images. Nonetheless, less attention has been paid to GANs in unsupervised disentanglement learning. Accordingly, we propose OOGAN, a novel framework based on GANs that can explicitly disentangle while generating high-quality images. The framework's components can readily be adopted to other GAN models.

Unlike in VAEs, where a latent vector has to be inferred, in GANs, noise is actively sampled as the latent vector during training. We exploit this property to enable OOGAN to directly learn a disentangled latent vector, by means of one-hot vectors as latent representations to enforce exclusivity

and to encourage each dimension to capture different semantic features. This is achieved without sacrificing the continuous nature of the latent space through an alternating sampling procedure. We argue that our proposed OOGAN fully highlights the structural advantage of GANs over VAEs for disentanglement learning, which, to the best of our knowledge, has not been exploited before.

We achieve disentanglement in OOGAN through three contributions: 1) We propose an alternating *one-hot sampling* procedure for GANs to encourage greater disentanglement. 2) We adopt an orthogonal regularization on the model weights to better accompany our objective. 3) We identify a weakness in InfoGAN and related models with similar structure, which we summarize as the *compete and conflict issue*, and propose a model-structural change to resolve it. Moreover, we propose a compact and intuitive metric targeting the disentanglement of the generative part in the models. We present both quantitative and qualitative results along with further analysis of OOGAN, and compare its performance against VAEs and InfoGAN.

As proposed by Locatello et al. (2018), we hereby clarify that the main inductive bias in this paper comes from our model design. Since we assume that there is a latent vector that controls the attributes of the data, and the dimensions of this vector are mutually independent. We design our model in view of these assumptions.

2 Related Work

β-VAE-based models: In the settings of β-VAE and its variants (Higgins et al. 2017; Burgess et al. 2018; Chen et al. 2018; Kim and Mnih 2018; Esmaeili et al. 2018), a factorized posterior $p_\phi(\mathbf{z} \mid \mathbf{x})$ is learned such that each dimension of a sampled z_i is able to encode a disentangled representation of data $\mathbf{x}$. The fundamental objective that β-VAE tries to maximize (also known as the Evidence Lower-Bound Optimization) is:

$$\mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}, \beta) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x} \mid \mathbf{z})] - \beta D_{\mathrm{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) || p(\mathbf{z})), \tag{1}$$

where $\beta > 1$ is usually selected to place stronger emphasis on the KL term for a better disentanglement learning. Burgess et al. (2018) motivate the effect of β from an information-theoretical perspective, where the KL divergence term can be regarded as an upper bound that forces $q(z)$ to carry less information, thus becoming disentangled.

Follow-up research extends the explanation by deriving a Total Correlation from the KL term in the β-VAE objective, and highlights this TC term as the key factor to learning disentangled representations. Given a multi-dimensional continuous vector z, the TC quantifies the redundancy and dependency among each dimension z_i. It is formally defined as the KL divergence from the joint distribution $q(z_1, ..., z_n)$ to the independent distribution of $q(z_1)q(z_2)...q(z_n)$:

$$\mathcal{L}_{\mathrm{TC}} = D_{\mathrm{KL}}(q(z)||\hat{q}(z)), \tag{2}$$

where $\hat{q}(z) = \prod_{i=1}^{n} q(z_i)$. However, the TC term requires the evaluation of the density $q(z) = \mathbb{E}_{p(n)}[q(z|n)]$, which depends on the distribution of the entire dataset and usually is intractable. For the sake of a better optimization on the TC term, Lample et al. (2017) propose TC-VAE, which uses a minibatch-weighted sampling method to approximate TC. Kim and Mnih (2018) perform the same estimation using an auxiliary discriminator network in their Factor-VAE. Furthermore, Esmaeili et al. (2018) suggest a more generalized objective where the marginals $q(z_i)$ can be further decomposed into more TC terms, in case each $q(z_i)$ learns independent but entangled features, which leads to a hierarchically factorized VAE. Dupont (2018) leverage the Gumbel Max trick (Jang, Gu, and Poole 2017) to enable disentangled learning of discrete features for VAE.

GAN-based models: InfoGAN (Chen et al. 2016) reveals the potential of GANs (Goodfellow et al. 2014) in the field of unsupervised disentanglement learning. In a typical GAN setting, a generator G and a discriminator D are trained by playing an adversarial game formulated as:

$$\min_G \max_D \mathcal{L}_{\mathrm{GAN}}(D, G) = \mathbb{E}_{p(x)}[\log(D(x))] + \mathbb{E}_{p(z)}[\log(1 - D(G(z))]. \tag{3}$$

While this mini-max game guides G towards generating realistic x from noise z drawn from the isotropic Gaussian distribution, the variation of z often remains entangled. InfoGAN manages to make G learn a disentangled transformation from a latent code c, which is concatenated to z before being fed to G. InfoGAN achieves this by maximizing a Mutual Information (MI) lower-bound between c and the generated sample $x = G(z, c)$, where the MI $I(c, G(z, c))$ can be calculated directly by matching c to $\hat{c} = Q(G(z, c))$, where Q is an auxiliary network that seeks to predict the sampled latent vector from x. In practice, Q shares most weights with D. However, such a lower-bound constraint only ensures c gains control over the generation process, but cannot guarantee any disentanglement as c increases its dimensionality, because this lower-bound does not encourage any independence across each dimension of c.

A more recent GAN-based disentanglement work is the Information-Bottleneck-GAN (Jeon, Lee, and Kim 2019). However, it fails to take advantage of the GAN structure, instead trying to implicitly minimize the TC in the same way as β-VAE. The method requires an extra network that encodes noise z into to a control vector c and lets the original G and D play the decoder's role to reconstruct z. This severely hurts the generation quality, since G starts the generation from c, which has a much lower dimensionality than z, and the increased network modules and loss objectives make the training scheme tedious and less likely to find the proper hyper-parameters that allow the model to converge.

3 Proposed Method

Our approach accomplishes both the task of disentangled feature extraction and human-controllable data generation in an unsupervised setting within the GAN framework. We define our problem as follows: For a continuous control vector c sampled from $\mathsf{uniform}(0, 1)$, we wish our generator G to be disentangled such that each dimension in c solely controls one feature of the generated data $x = G(c, z)$ (z is the noise

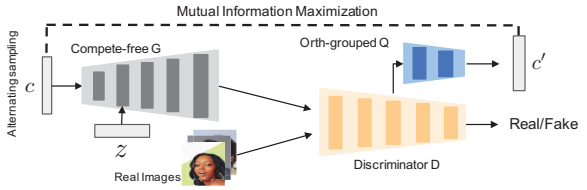


Figure 1: OOGAN makes minimal changes upon a basic GAN. c denotes the continuous control vector, z is the noise vector, c' is the feature representation of fake images.

vector), and our feature extractor Q (mostly the discriminator D with a few layers on top that gives vector outputs) is able to emit a feature representation c', given x, that is disentangled in the same way as c.

Our model is illustrated in Figure 1. Similar to the design of InfoGAN, we let the feature extractor Q be a sub-module that shares weights with the discriminator D. Q takes the feature map of a generated image $G(c, z)$ as input and tries to predict the control vector c used by the generator G. We describe the three components of our OOGAN framework in the following sub-sections.

Alternating Continuous and One-hot Sampling

Previous methods of minimizing TC to achieve disentanglement have two limitations. First, due to the intractability, extra network modules and objectives have to be invoked to approximate TC, which leads to undesired hyper-parameter tuning, a non-trivial training regime, and a high computational overhead. Second, to optimize the derived TC objectives in VAE-based models, the data generation quality is sacrificed (Kim and Mnih 2018; Chen et al. 2018), and can hardly perform well on higher-resolution image data. In contrast, in the GAN setting, the latent vector is sampled instead of inferred as in VAEs. This motivates us to approach disentanglement by deliberately sampling latent vectors that possess the property of inter-dimensional independence and training the networks using these sampled vectors.

To this end, we propose an alternating continuous-discrete sampling procedure: we alternate between sampling continuous c from $\mathsf{uniform}(0, 1)$ (as typically done in InfoGAN) and sampling c as one-hot vectors. The one-hot vector c implies that the generated image should only exhibit one feature, and, ideally, the prediction c' from Q should also be a one-hot vector. On both the G and Q sides, any presence of other features should be penalized, while alternating with continuous uniform sampling is necessary to ensure the continuity of the representation. Interestingly, such a one-hot sampling resembles a classification task. Therefore, we can jointly train Q and G directly via a cross-entropy loss. In such a process, G is trained to generate images that possess the specified features and avoid retaining any other features, while Q is trained to summarize the highlighted feature only in one dimension and refrain from spreading the feature representation into multiple dimensions.

Note that we treat c as a continuous vector in the entire training process, and the alternating one-hot sampling can be seen as a regularizor for G and Q. When we sample c from $\mathsf{uniform}(0, 1)$ as in InfoGAN, we ensure the correlation between c and x remains. Furthermore, when we interleave that with one-hot samples, the process can be interpreted as getting the extremely typical samples (those samples that lie on the boundary of the uniform distribution) for the model. We argue that sampling data at the distribution boundaries makes the model pay more attention to these boundaries, yielding a clearer distribution shape highlighting the semantics of these boundary factors. These typical samples are vital for the model to learn inter-dimension exclusivity, as a one-hot c regularizes G to generate images with only one factor and Q to only capture this one factor.

In other words, alternating one-hot and uniform sampling results in a more desirable prior distribution for disentangling GANs, which provides more typical samples on the margin than a single uniform distribution. Such an alternating procedure, which injects the categorical sampling (i.e., the one-hot sampling) into a continuous c, makes it possible that c gains continuous control over the generation process while simultaneously paying more attention to those typical examples, and therefore achieves better disentanglement.

Formally, our complete objective for OOGAN is:

$$\begin{aligned}\min_G \max_D \mathcal{L}_{\text{OOGAN}}(D, G) = \mathcal{L}_{\text{GAN}}(D, G) + \\ \lambda I(c_{\text{continuous}}, G(c_{\text{continuous}}, z)) + \\ \gamma \mathcal{L}_{\text{Cross-Entropy}}(Q(G(c_{\text{one-hot(d)}}, z)), c_{\text{one-hot(d)}}).\end{aligned} \tag{4}$$

Despite the lack of any TC terms in our objective, the one-hot sampling still ensures that we have a well-disentangled feature extractor Q and generator G that learn features with no overlap between each dimension in c, without any approximations and extra network modules involved.

Compete-Free Generator

InfoGAN (Chen et al. 2016) and many conditional-GAN variants leverage an auxiliary vector c that is concatenated with noise z before being fed into G, with the expectation that c carries the human-controllable information. From a size perspective, the dimensionality of z is usually much more significant than that of c (z typically has around a hundred dimensions, while c has in the order of 10). Intuitively, c will have much less impact in the generation process. With the objective of unsupervised disentanglement learning, a large portion of influence z takes in the generation process is undesirable, which we refer to as the ***competing and conflicting issue***.

Usually, a disentangled feature learned by c can also be entangled in z. During the training process, if c with c_i holding a high signal on a certain feature is paired with some z with many dimensions holding the same feature with a conflicting signal, this signal, entangled in z, will easily overpower c. Thus, the generated images will not present c_i's signal. Such a conflict will discourage c from mastering the learned feature and cause it to stray away to some easier-to-achieve but less distinct features. An example is shown in Figure 2. More discussion can be found in the experiments section.

To avoid the aforementioned *competing and conflicting issue*, we propose a new **Compete-Free Design** of the gen-

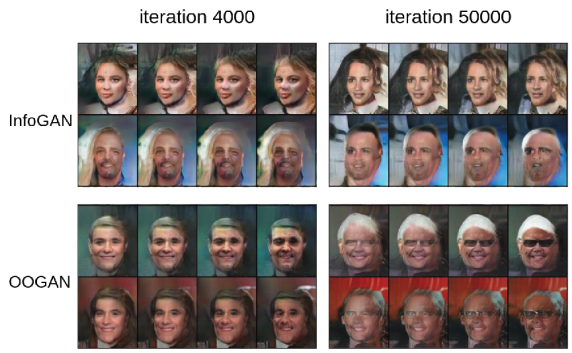


Figure 2: Latent traversals trained on CelebA to showcase the *competing and conflicting issue*. The images are from the same set of (z, c) on one fixed dimension of c after different training iterations. We observe that InfoGAN begins to capture what appears to be a "wearing glasses" feature at a very early stage, but discards it during training in all dimensions of c. In contrast, when OOGAN begins to capture this feature, it consistently masters it in the end.

erator's input block, which switches the role between c and z by letting c control the fundamental content even when the dimensionality of c is low, and ensures that z has limited influence in the generation process.

To start with, we project the low-dimensional control vector c into a multi-channel 4×4 feature map by a convTransposed layer. Then, we add this feature map to a learned constant tensor with the same dimensionality.

The weights for the constant tensor are randomly initialized before training and are trained via back-propagation just like all other model weights. This learned constant can be regarded as an additive bias that is learned from the dataset, and is necessary since it is responsible for representing the features that are not captured in c. Ideally, when given a c with all zeros as input, this constant should let the generator output the most "neutral" x. In our experiments, we find such constant important for a more stabilized learning process. It makes OOGAN faster to converge to the disentangled factors. Intuitively, one can imagine this constant as placing an anchor at the center of the target distribution, such that all latent factors can expand in different directions. This behavior encourages the model to focus on learning the correlation between c and the generated images. Without this constant, OOGAN will still work but will be slower to converge for c.

To encourage the variance and complement the details for a higher-quality generation, the traditional noise z is still taken into the generator, but only after the 8×8 feature map level. To prevent z from causing the *competing and conflicting issue*, we leverage an attention mask generated from c on the features from z, which means that only the approved part of z by c can join in the generation process. Different layers in CNNs have been studied extensively (Karras, Laine, and Aila 2019), where the first few layers tend to generate fundamental compositions, and higher layers only refine the details. So our design makes c more natural to control the

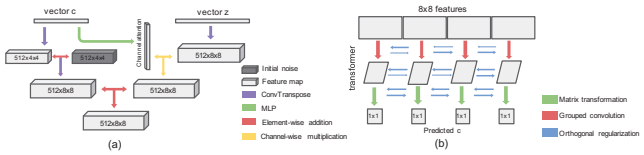


Figure 3: Model structures: (a) Input block of the competefree G. (b) Orthogonal-regularized grouped Q

key generative factors without the interference of z. The design details are illustrated in Figure 3-(a).

Our generator design resembles the one proposed for StyleGAN (Karras, Laine, and Aila 2019), as we both base the generation on a fixed multi-dimensional feature map instead of an input vector z, and take z as input only in later layers. As claimed by Karras *et al.*, such a design leads to a better separation in the data attributes and a more linear interpolation along latent factors. However, both the motivation and structure details are different. The *disentanglement* we study here is a more strict term than what Karras *et al.* used. The fundamental difference is that the fixed weights in our proposed model only serve as a supportive bias, and will be directly changed by c, while in StyleGAN the fixed weights are solely used to start the image generation process.

Orthogonal Regularized & Grouped Feature Extractor

To learn a disentangled representation, we propose a new structure of Q that uses grouped convolutions (Krizhevsky, Sutskever, and Hinton 2012; Zhang et al. 2018) instead of traditional fully connected ones, with an orthogonal regularization on the weights among every convolution kernel. The intuition is, since we hope that Q will be a highly disentangled feature extractor, a fully connected (FC) design is not favorable, since, in a FC convolution, each feature prediction has to take into consideration all the feature maps from the previous layer. A grouped convolution, on the other hand, can focus its decision making on a much smaller group of previous features, and may thus be less distracted by potentially irrelevant features.

To make sure that each group is indeed attending to different features, we impose an additional loss function on the weights of the convolutional layers to enforce the orthogonality between different kernels. Weight orthogonality in DNNs has been studied (Brock et al. 2016; Huang et al. 2018; Bansal, Chen, and Wang 2018). However, these studies each focused on different tasks, and none of them revealed the potential for disentanglement learning.

The orthogonal regularization we use is straightforward: during each forward pass of the OOGAN, compute and minimize the cosine similarity between every convolutional kernel. With grouped feature extraction and orthogonal regularization, Q structurally more easily captures diversified features in each dimension. Note that the group design is not only applicable to convolutional layers but also to grouped linear layers or other weights indicated as "transformer" in Figure 3-(b). Similarly, the orthogonal regularization can be applied on weights of all these grouped layers.

4 Perceptual Diversity Metric

Quantitative metrics for the disentanglement are mostly proposed in VAE-based works and for simulated toy datasets with available ground truth information. Higgins et al. (2017) suggest training a low-capacity linear classifier on the obtained latent representations of the simulated data from the trained encoder, and report the error rate of the classifier as the disentanglement score of the generative model. Kim and Mnih (2018) argue that the introduction of an extra classifier could lead to undesirable uncertainties due to the increased hyper-parameters to tune. Thus, they favor a majority-vote classifier that can be obtained more directly. We concur with Kim and Mnih (2018) in arguing that, to the best of our knowledge, there is no convincing metric for disentanglement on a dataset for which no ground truth latent factors are provided. Therefore, we propose a method that is capable of relatively evaluating partial properties of a disentangling model when certain conditions are satisfied.

Our intuition is that if a generative model is well-disentangled, then varying each dimension of the controlling vector c should yield different feature changes of the generated data x. Suppose the feasible value range for c is $[a, b]$, and for a pair of (c^o, i, j) where c^o is a uniformly sampled vector and i and j are two randomly selected indices, we get c^i by setting $c^o[i] = b$ and $c^o[j] = a$, and c^j by setting $c^o[j] = b$ and $c^o[i] = a$. Given the fact that i and j each control different factors, we expect $x^i = G(c^i)$ and $x^j = G(c^j)$ to be different. Therefore, we can use a pre-trained VGG (Simonyan and Zisserman 2014) model V to extract the feature map of x^i and x^j, and report their L_1 distance as the disentanglement score, with a higher L_1 distance indicating dimensions i and j are more independent. The final score of this proposed *perceptual diversity metric* will be the average score of many samples of paired (c^o, i, j).

We argue that such a metric can adequately reflect the separability and diversity of the learned factors, especially when used for comparing similarly structured models on high-resolution datasets, where higher diversity should already be considered better, and on datasets in which latent factors are known to control a good amount of visual differences. As shown in Figure 4, the proposed metric can efficiently capture the disentangle performance in terms of how diversified each dimension is in c.

5 Experiments

We conduct quantitative and qualitative experiments to demonstrate the advantages of our method on several datasets. First, we perform quantitative experiments on the dSprites datasets (Matthey et al. 2017) following the metric proposed by Kim and Mnih (2018). After that, we show the superiority of OOGAN in generating high-quality images while maintaining competitive disentanglement compared to VAE-based models on CelebA (Liu et al. 2015) and 3D-chair (Aubry et al. 2014) data. Based on the disentangling benchmark guidance from Eastwood and Williams (2018), we also present an elaborated learned-factor identification experiment to showcase the effectiveness of OOGAN and validate our compete-and-conflicting issue observations.

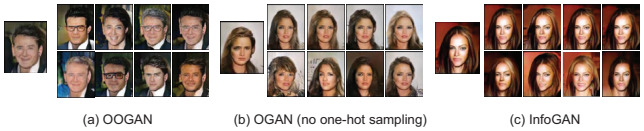


Figure 4: Generated images for CelebA: In each group, the left-most image is generated from a randomly sampled c, and the following ones are generated by changing the value of each dimension in c to 1. (a) OOGAN exhibits greater visual differences among each dimension, reflecting its ability to learn diverse latent factors. (b) Without the proposed one-hot sampling, OOGAN still manages to learn some distinguishable features, reflecting the advantage of its structural design. (c) The 4 top right images show that the learned features for an InfoGAN have a large overlap across the latent dimensions in c, lacking proper disentanglement.

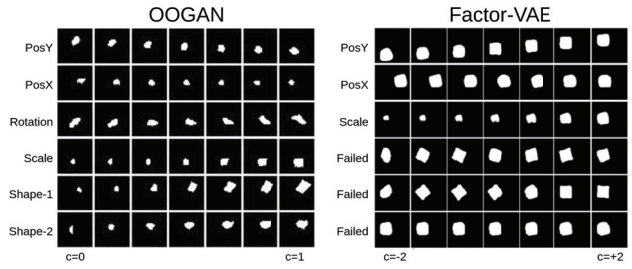


Figure 5: Latent traversals on dSprites

Finally, we conduct an ablation study on the proposed components in OOGAN with our metric.

Hardware and training conditions: We perform all the experiments on one NVIDIA RTX 2080Ti GPU, on which all the models can be well trained within (up to) 10 hours. All the code to reproduce our experiments is available on GitHub, and training configurations can be found there.

Quantitative results on dSprites: Several quantitative metrics have been proposed on the dSprites dataset (Higgins et al. 2017; Kim and Mnih 2018; Eastwood and Williams 2018; Chen et al. 2018). While these metrics achieve a thorough evaluation of the disentanglement abilities of the feature-extractor (i.e., the encoder in VAE and Q in GANs), they pay no attention to the generative part of the models. Therefore, we only select Kim and Mnih's metric for its intuitiveness and simplicity to demonstrate our model's competitiveness on the feature extractor's end.

For all the models, we follow the same setup as Kim and Mnih (2018) and Jeon, Lee, and Kim (2019). Due to the simplicity of the dataset, we train all the GAN models with the "instance noise technique" introduced by Sønderby et al. (2017) to get stable and good quality results.

As can be seen from Table 1 and Figure 5, our proposed OOGAN genuinely does a better job on both the feature extractor and generator parts. While Factor-VAE is only able to disentangle three out of the five ground truth factors effectively, OOGAN retrieves all the generative factors and manages to put the variables of the discrete factor "shape" into

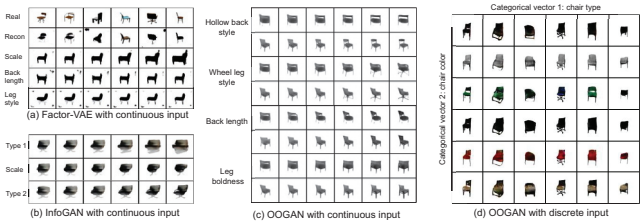


Figure 6: Latent traversals on 3D Chair

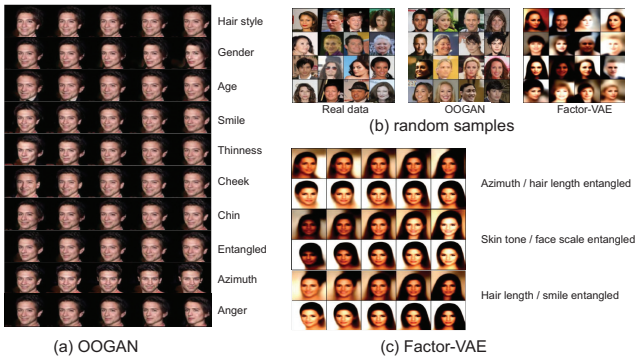


Figure 7: Latent traversals for model trained on CelebA

different dimensions. Additionally, we would like to highlight the robustness of our model, where varying the hyperparameters of λ (1 to 5) and γ (0.2 to 2) in our loss function always yields consistent performance.

Qualitative results on 3D Chairs: On the 3D Chairs data, we use 64×64 RGB images with batch size 64 for all training runs. To demonstrate the robustness and performance of OOGAN in generating higher-quality images and potentially learning more latent factors, we consider dimensionalities of c of up to 16, where previous works only experiment on a smaller dimensionality such as 6.

In Factor-VAE, when we increase the dimensionality of c from 6 to 16, it struggles to disentangle at a similar quality, and the reconstruction ability is severely sacrificed. In contrast, our model is not affected by an increase of the dimensionality, and apart from learning somewhat more obvious features such as scale and azimuth, our model also discovers several exciting features that have never been reported in previous work. For example, Figure 6-(c) shows a linear transformation of different back styles and leg thickness of the chairs, and Figure 6-(d) shows that our model successfully disentangles discrete features such as "color" and "chair type" without any additional tweaks and tricks, for which additional tweaks such as various approximation approaches would have to be incorporated in a VAE approach.

Disentangling at a higher resolution on CelebA: We consider OOGAN as a suite of three modules that can be plugged into any GAN frameworks, and it is orthogonal to other disentanglement approaches based on GANs such as IBGAN. In other words, it can be incorporated into other methods and inherits the breakthroughs made in GANs (Zhang et al. 2019; Miyato et al. 2018; Karras, Laine, and Aila 2019). Therefore, we focus on demonstrating the advantages OOGAN has over VAE-based models in qualitative experiments. The comparison with InfoGAN will be presented as quantitative results in an ablation study.

Table 1: Disentanglement using Kim *et al.*'s metric

Model	Score
β-VAE	0.63 $\pm$0.033
Factor-VAE	0.73 $\pm$0.112
InfoGAN	0.59 $\pm$0.078
IB-GAN	0.80 $\pm$0.062
OOGAN	0.81 $\pm$0.077

Table 2: Disentanglement using Perceptual Diversity metric

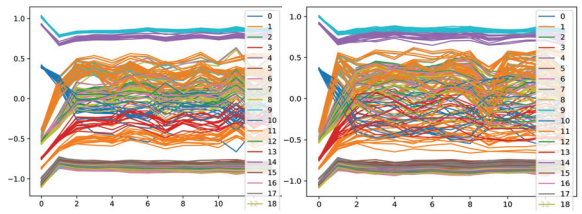
Model	Score	Cos-simil. in Q
InfoGAN	2.39 $\pm$0.03	0.21 $\pm$ 0.01
OOGAN w/o One-hot	2.44 $\pm$0.05	0.09 $\pm$ 0.03
OOGAN w/o Ortho-reg	2.65 $\pm$0.05	0.21 $\pm$ 0.01
OOGAN w/o Compt-free G	2.69 $\pm$0.03	0.09 $\pm$ 0.03
OOGAN	2.77 $\pm$0.06	0.09 $\pm$ 0.03

On the CelebA dataset, while previous work operates at a resolution of 64×64, we train all the models at a resolution of 256×256 to showcase the advantage of OOGAN and expose a shortcoming of the VAE-based models. Figure 4-(a) shows the images trained in a plain DCGAN manner (Radford, Metz, and Chintala 2015) and Figure 7-(a) shows the images trained in a progressively up-scaling manner (Karras et al. 2018), demonstrating a strong ability to disentangle while maintaining a high image quality. On the other hand, VAE-based models deteriorate when reconstructing high-contrast images, and are unable to maintain the same disentanglement performance as the resolution increases. Thanks to the detail richness of the generated images, OOGAN can discover more interesting facial features such as "chin" and "cheek", which no VAE-based models have achieved.

Learned attribute analysis: To provide a more transparent breakdown on what is learned, we train 40 binary classifiers on the 40 provided visual attributes from the CelebA dataset, each only predicting one attribute. Then we use these classifiers to monitor the generated images across the training iterations of InfoGAN and OOGAN.

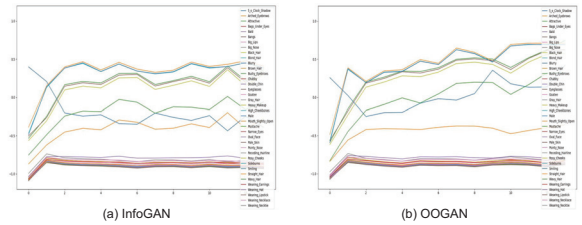
A shown in Figure 8, there are 40 different colors of lines, each color representing an attribute, with 16 lines per color representing the 16 dimensions in c. In terms of sampling c, we set one dimension's value to 1 and sample the remaining dimensions from $\mathsf{uniform}(0, 1)$, repeating the same operation for all dimensions. For InfoGAN, the lines with the same color stick together and have the same tendency to change, which means different dimensions in c are learning similar factors. In contrast, for OOGAN, we observe that lines in the same color (same attribute) develop differently during the iterations. Moreover, at each iteration, the lines in the same color get different prediction scores, implying that different dimensions in c are learning different factors.

We then average the curves of each attribute into one to show the prediction score for each ground truth attribute in Figure 9, which confirms the **competing and conflicting is-**



(a) InfoGAN (b) OOGAN

Figure 8: Binary classification of each labeled attributes for each dimension on celebA



(a) InfoGAN (b) OOGAN

Figure 9: Binary classification of each labeled attributes over all dimensions on celebA

sue. If a prediction score is constantly increasing, this means that c is improving its ability to represent the respective attribute. For InfoGAN, the attribute predictions show rising and falling fluctuations, a sign of an unstable learning process. In contrast, OOGAN steadily increases the prediction score for some visual attributes.

Ablation studies: Based on the plain InfoGAN setting, we conduct ablation studies on the effectiveness of our proposed three modules quantitatively using our proposed metric, with InfoGAN as the baseline. The experiments are conducted only on learning continuous factors, as InfoGAN already performs well in disentangling categorical latent variables.

There are two types of Q we can choose from when training OOGAN or InfoGAN. A *deterministic* Q will try to directly output c', which is considered as a reconstruction of c; and a *probabilistic* Q will assume that each dimension of c_i is from a Gaussian and try to output the mean and standard deviation of that distribution given different input images. To optimize a deterministic Q, we can directly minimize the L1 loss between predicted c' and the input c, and to optimize the probabilistic Q, we can minimize the negative log-likelihood given the sampled c and predicted μ and σ. In both cases, our proposed one-hot sampling trick can participate in the optimization directly, where for the probabilistic Q, we just minimize the cross-entropy between μ and c.

When the dimensionality of c goes beyond 6, InfoGAN fails to disentangle them, which is also confirmed in previous work (Kim and Mnih 2018; Jeon, Lee, and Kim 2019). As shown in Figure 4, most dimensions in InfoGAN produce similar images, as many features remain entangled. In the meantime, OOGAN has a better tendency to learn disentangled representations thanks to its structural design, and a direct objective to learn independent features driven by the proposed alternating one-hot sampling.

For the proposed perceptual diversity metric, we fine-

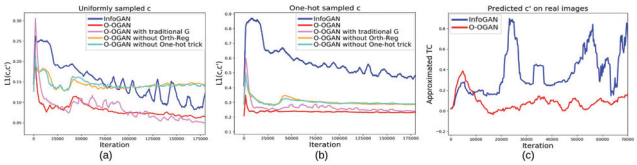


Figure 10: (a)&(b): L1 losses between sampled c and predicted c'. (c) TC estimation during training

tune the VGG model on the CelebA dataset with the provided 40 facial attributes, to make it more sensitive to the visual attributes. As shown in Table 2, the one-hot sampling makes the most substantial contribution, while orthogonal-regularized Q and compete-free generator also provide significant improvements. The averaged cosine similarity among the weights is effectively minimized with the proposed orthogonal regularization. In Figures 10-(a) and 10-(b), we train the models with a deterministic Q that directly attempts to reconstruct c, and plot the L1 distance between the sampled c and predicted c' (the L1 distance for one-hot c is not used as an objective loss to train InfoGAN). Note how InfoGAN's L1 loss is similarly minimized when c is uniformly sampled, but struggles to decrease when c is one-hot, which means that the output c' of InfoGAN is highly correlated (there are correlated latent factors encoded into multiple dimensions, implying poor disentanglement), while OOGAN's c' is not. In Figure 10-(c), we train the models with probabilistic Q and estimate TC following the method from Chen et al. (2018). The TC from InfoGAN remains high, while OOGAN can maintain a low TC consistently, which shows the effectiveness of our method.

6 Conclusion

We propose a robust framework that disentangles even high-resolution images with high generation quality. Our one-hot sampling highlights the structural advantage of GANs for easy manipulation on the input distribution that can lead to disentangled representation learning, while the architectural design provides a new perspective on GAN designs. Instead of tweaking the loss functions (designing a new loss, adjusting loss weights, which are highly unstable and inconsistent across datasets), we show that sampling noise from multiple distributions to achieve disentanglement and interpretability is robust and straightforward. It leads to a promising new direction of how to train GANs, opening up substantial avenues for future research, e.g., choosing what distribution to sample from, allocating alternating ratios, etc. The impact of this goes beyond disentangling, as future research can also be conducted on model interpretability and human-controllable data generation. In the future, we plan to explore more dynamic and fluent sampling methods that can be integrated into the GAN framework for better performance, and we will attempt to validate the benefits of these sampling methods theoretically.

Acknowledgment This work is partially supported by NSFUSA award 1409683.

References

Alemi, A.; Poole, B.; Fischer, I.; Dillon, J.; Saurus, R. A.; and Murphy, K. 2018. An information-theoretic analysis of deep latent-variable models. *https://openreview.net/forum?id=H1rRWl-Cb*.

Aubry, M.; Maturana, D.; Efros, A.; Russell, B.; and Sivic, J. 2014. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In *CVPR*.

Bansal, N.; Chen, X.; and Wang, Z. 2018. Can we gain more from orthogonality regularizations in training deep networks? In *NeurIPS*.

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation learning: A review and new perspectives. *T-PAMI* 35(8).

Brock, A.; Lim, T.; Ritchie, J. M.; and Weston, N. 2016. Neural photo editing with introspective adversarial networks. In *ICLR*.

Burgess, C. P.; Higgins, I.; Pal, A.; Matthey, L.; Watters, N.; Desjardins, G.; and Lerchner, A. 2018. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*.

Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; and Abbeel, P. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NeurIPS*.

Chen, T. Q.; Li, X.; Grosse, R. B.; and Duvenaud, D. K. 2018. Isolating sources of disentanglement in variational autoencoders. In *NeurIPS*.

Dupont, E. 2018. Learning disentangled joint continuous and discrete representations. In *NeurIPS*.

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

Elgammal, A.; Liu, B.; Kim, D.; Elhoseiny, M.; and Mazzone, M. 2018. The shape of art history in the eyes of the machine. In *AAAI*.

Elhoseiny, M.; Zhu, Y.; Zhang, H.; and Elgammal, A. 2017. Link the head to the "beak": Zero shot learning from noisy text description at part precision. In *CVPR*.

Esmaeili, B.; Wu, H.; Jain, S.; Bozkurt, A.; Siddharth, N.; Paige, B.; Brooks, D. H.; Dy, J.; and van de Meent, J.-W. 2018. Structured disentangled representations. In *AISTATS*.

Esmaeili, B.; Wu, H.; Jain, S.; Bozkurt, A.; Siddharth, N.; Paige, B.; Brooks, D. H.; Dy, J.; and Meent, J.-W. 2019. Structured disentangled representations. In *ICAIS*.

Gabrié, M.; Manoel, A.; Luneau, C.; Macris, N.; Krzakala, F.; Zdeborová, L.; et al. 2018. Entropy and mutual information in models of deep neural networks. In *NeurIPS*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NeurIPS*.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*.

Huang, L.; Liu, X.; Lang, B.; Yu, A. W.; Wang, Y.; and Li, B. 2018. Orthogonal weight normalization: Solution to optimization over multiple dependent stiefel manifolds in deep neural networks. In *AAAI*.

Jang, E.; Gu, S.; and Poole, B. 2017. Categorical reparameterization with gumbel-softmax. In *ICLR*.

Jeon, I.; Lee, W.; and Kim, G. 2019. IB-GAN: Disentangled representation learning with information bottleneck GAN. *https://openreview.net/forum?id=ryljV2A5KX*.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *CVPR*.

Kim, H., and Mnih, A. 2018. Disentangling by factorising. In *ICML*.

Kim, D.; Liu, B.; Elgammal, A.; and Mazzone, M. 2018. Finding principal semantics of style in art. In *ICSC*.

Kingma, D. P., and Welling, M. 2014. Auto-encoding variational bayes. In *ICLR*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NeurIPS*.

Lample, G.; Zeghidour, N.; Usunier, N.; Bordes, A.; DENOYER, L.; and Ranzato, M. A. 2017. Fader networks: Manipulating images by sliding attributes. In *NeurIPS*.

Lin, Z.; Thekumparampil, K. K.; Fanti, G.; and Oh, S. 2019. InfoGAN-CR: Disentangling generative adversarial networks with contrastive regularizers. *arXiv preprint arXiv:1906.06034*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *ICCV*.

Locatello, F.; Bauer, S.; Lucic, M.; Gelly, S.; Scholkopf, B.; and Bachem, O. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *ICML*.

Louizos, C.; Ullrich, K.; and Welling, M. 2017. Bayesian compression for deep learning. In *NeurIPS*.

Matthey, L.; Higgins, I.; Hassabis, D.; and Lerchner, A. 2017. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/.

Miyato, T.; Kataoka, T.; Koyama, M.; and Yoshida, Y. 2018. Spectral normalization for generative adversarial networks. In *ICLR*.

Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*.

Saxe, A. M.; Bansal, Y.; Dapello, J.; Advani, M.; Kolchinsky, A.; Tracey, B. D.; and Cox, D. D. 2018. On the information bottleneck theory of deep learning. In *ICLR*.

Shannon, C. E. 1948. A mathematical theory of communication. *Bell system technical journal* 27(3).

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Sønderby, C. K.; Caballero, J.; Theis, L.; Shi, W.; and Huszár, F. 2017. Amortised map inference for image super-resolution. In *ICLR*.

Watanabe, S. 1960. Information theoretical analysis of multivariate correlation. *IBM Journal of research and development* 4(1).

Zhang, X.; Zhou, X.; Lin, M.; and Sun, J. 2018. Shufflenet: an extremely efficient convolutional neural network for mobile devices. In *CVPR*.

Zhang, H.; Goodfellow, I.; Metaxas, D.; and Odena, A. 2019. Self-attention generative adversarial networks. In *ICML*.

Zhu, Y.; Elhoseiny, M.; Liu, B.; Peng, X.; and Elgammal, A. 2018. A generative adversarial approach for zero-shot learning from noisy texts. In *CVPR*.

Zhu, Y.; Xie, J.; Liu, B.; and Elgammal, A. 2019a. Learning feature-to-feature translator by alternating back-propagation for zero-shot learning. In *ICCV*.

Zhu, Y.; Xie, J.; Tang, Z.; Peng, X.; and Elgammal, A. 2019b. Semantic-guided multi-attention localization for zero-shot learning. In *NeurIPS*.