RETHINKING CONTENT AND STYLE: DIGGING BIAS FOR UNSUPERVISED DISENTANGLEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Content and style disentanglement intends to decompose the underlying explanatory factors of objects into two different latent spaces. Content characterizes the geometry structure attributes shared across instances, e.g., pose and shape, while style models instance-specific properties, e.g., identity and texture. In this paper, we provide a novel formulation of the content and style disentanglement problem and demystify the inductive bias contained in this framework for unsupervised disentanglement. Based on our formulation, the content embeddings are forced to share a common distribution, the style embeddings encoding instance specific feature are used to customize the shared distribution to match the real data distribution. The experiments on several popular datasets demonstrate that our method achieves state of the art disentanglement when compared to other unsupervised approaches, and also comparable or even better results than supervised methods. Thanks to our disentangled representation, we generate multi-views from a single view for 3D reconstruction, which demonstrates that the pose and identity are well disentangled.

1 INTRODUCTION

Disentanglement task aims to recover the the underlying explanatory factors of nature images into different dimensions of latent space, and provide an informative representation for the tasks like image translation (Wu et al., 2019b; Kotovenko et al., 2019), domain adaptation (Li et al., 2019; Zou et al., 2020) and geometric attributes extraction Wu et al. (2019c); Xing et al. (2019), et al. Some recent works (Denton & Birodkar, 2017; Jha et al., 2018; Bouchacourt et al., 2018; Szabó et al., 2018) rely on supervision, e.g., group observation (Gabbay & Hoshen, 2020a) which is hard to obtain for real data. For unsupervised methods (Kim & Mnih, 2018; Higgins et al., 2017; Burgess et al., 2018b; Kumar et al., 2017), Locatello et al. (2019) shown that this task is fundamentally impossible, but can be achieved by introducing inductive bias on both model and data.

In this paper we focus on Content ans Style (C-S) disentanglement, which decomposes the image embeddings into two separate space. Content includes the factors shared across the whole object category, like the pose and shape of object. Style reflects instance-specific attributes, like the identity and texture attributes. Take the human face dataset for example, multiviews of the same face share the same style embeddings, but have different content embeddings to represent different pose.

To achieve unsupervised C-S disentanglement, the inductive bias is necessary. We rethink this task by formulating the problem, which provides a novel view to better understand the nature of the task. Based on this formulation, we propose a design to introduce inductive bias. The key is to assign different and independent role to content and style when approximating real data distributions: we force the content embeddings to share a common distribution and the style embeddings are used to scale and shift the common distribution to match target image distribution.

The embeddings and the parameters of the generator which maps embeddings to images is optimized by latent optimization (Bojanowski et al., 2018; Gabbay & Hoshen, 2020a). Please note that we only use the image reconstruction loss as the supervision, no extra labeling are needed. As the content and style act independent role when modeling the data, after the optimization, they will be disentangled to encode the shared and specific feature respectively.

The contributions of our work are summarized as follows.

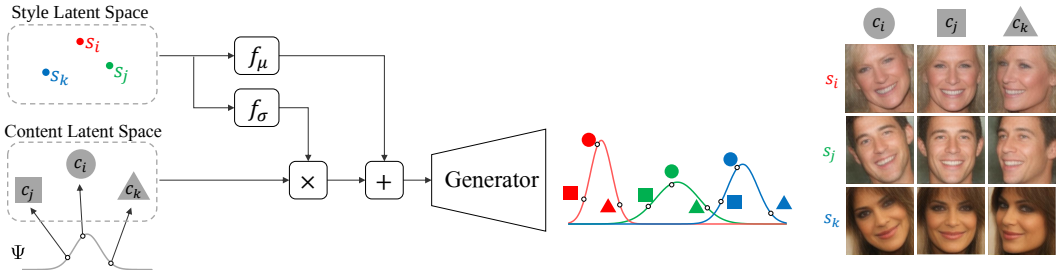


Figure 1: Overview of our framework. c_i, c_j, c_k labelled with different shapes are embeddings sampled from a shared distribution Ψ. s_i, s_j, s_k labelled with different colors are embeddings from style latent space. f_σ and f_μ are two MLPs predict the statics parameters to scale and shift Ψ respectively to approximates the target image distributions via a Generator. The 3×3 images shows that the pose of faces sampled from same positions of Ψ (with same shape) are aligned very well.

- We provide a novel formulation for C-S disentanglement problem to better understand the nature of this task.
- We achieve unsupervised C-S disentanglement by introducing inductive bias: assign different independent role to content and style when modeling the real data distributions.
- The experiments on several popular datasets demonstrates that our method achieves state of the art unsupervised C-S disentanglement, and comparable or even better results than supervised methods. We also apply C-S disengagement to a new task: single view 3D reconstruction task Wu et al. (2020).

2 Related Work

Disentangled Representation Learning. A disentangled representation can be defined as one where individual latent units are sensitive to changes in individual generative factors. There have been a lot of studies on disentangled representation learning: β-VAE (Higgins et al., 2017), Anneal VAE (Burgess et al., 2018b), DIPVAE (Kumar et al., 2017), FactorVAE (Kim & Mnih, 2018), BetaTCVAE (Chen et al., 2018). These models learn disentangled factors by factorising aggregated posterior. The learned factors can be divided into two categories, one is content-related, the other is style-related. Using these models is a simple direct method to unsupervisedly learn style content disentangled representation. However, Locatello et al. (2019) proved that unsupervised disentanglement is impossible without inductive bias on both model and data, therefore, these models are currently unable to obtain a promising disentangled representation. Motivated by Locatello et al. (2019), we leverage model inductive bias to disentangle style and content in an unsupervised manner. In addition, motivated by Anneal VAE (Burgess et al., 2018b), we enlarge such model inductive bias by information bottleneck.

Another branch of method focuses on dividing latent variable into two subspaces relying on class supervision. Mathieu et al. (2016) and Szabó et al. (2018) combine the adversarial constraint and auto-encoders. Variational autoencoder (Kingma & Welling, 2014) is also used with non-adversarial constraints, such as: cycle consistency (Jha et al., 2018) and evidence accumulation (Bouchacourt et al., 2018). Furthermore, latent optimization is shown to be superior to amortized inference (Gabbay & Hoshen, 2020a). Different from the above works, Wu et al. (2019c) propose to use variational U-Net with structure learning in an unsupervised manner. However, this method is limited to unsupervised learning of 2D landmark. Our method leverages network structure bias to directly learning two disentangled subspaces.

Style-Based Generator. In style transfer, feature statistics have been found to carry the style information of an image (Gatys et al., 2016; Li & Wand, 2016; Li et al., 2017). Motivated by this, Huang & Belongie (2017) make an extension to instance normalization, namely adaptive instance normalization (AdaIN). Karras et al. (2019) proposes to use AdaIN in a generator architecture for building StyleGAN. StyleALAE (Pidhorskyi et al., 2020) combines two autoencoders to further improve attributes separation of StyleGAN in the AdaIN path. To control image synthesis, pre-trained semantic predictors are used to conduct StyleGAN during inference (Shen et al., 2020; Esser et al.,

2020). Recently, Alharbi & Wonka (2020) demonstrate the disentanglement in the first layer of the generator network. While the above methods mainly focus on controllable image synthesis, we elegantly utilize style-based generator to disentangle representation by exploring the difference between its two paths.

3 METHOD

In this section, we first formulate the C-S disentangle problem and explain the inductive bias we brought into the data and model. Then we perform the ablation study to figure out how the C-S disentangled. Finally we introduce some important techniques of our method.

3.1 PROBLEM FORMULATION

The inductive bias on the dataset and model is proved to be the key for unsupervised disentanglement (Locatello et al., 2019). For the given dataset $D = \{I_i\}_{i=1}^N$, we assume each image I_i is sampled from a unknown distribution $P_i(x|y)$, where y is a shared conditional variable, which indicates the distributions hold some similar properties carried by y. This assumption serves as the inductive bias on dataset.

To approximate the target distribution $P_i(x|y)$, we define a distribution $Q_\theta(x|c_i, s_i)$ parameterized by θ, where θ is the parameter of the generator G_θ that maps embeddings to images, c_i and s_i are the content and style embeddings of I_i respectively. To introduce the inductive bias into model, we assign different and independent role to content and style when formulating Q. We restrict $\{c_i\}_{i=1}^N$ are sampled from the same distribution Ψ to ensure that the content attributes are shared across all the images.

Inspired by the observation that feature mean and variance carry the style information (Gatys et al., 2016; Li & Wand, 2016; Li et al., 2017; Huang & Belongie, 2017), we use the style embeddings to provide the statics to scale and shift the shared distribution Ψ to match target distribution as

$$z_i = f_\sigma(s_i) \cdot c_i + f_\mu(s_i) \tag{1}$$

where f_σ and f_μ are two fully connected layers predicting mean and variance respectively.

Our optimization target is to minimize the KL distance between Q and P,

$$\min_{\theta, c_i, s_i} \sum_{i=1}^{N} KL(p_i(x|y)||Q_\theta(x|c_i, s_i)) \tag{2}$$

This is equivalent to minimizing the image reconstruction loss as

$$\min_{\theta, c_i, s_i} \sum_{i=1}^{N} \|I_i - G_\theta(z_i)\| \tag{3}$$

Gaussian distribution is shown to be general and proper to model the real data (Kingma & Welling, 2014; Kim & Mnih, 2018; Higgins et al., 2017) and normalizing the feature map of network to Gaussian is helpful for the training (Ioffe & Szegedy, 2015; Huang & Belongie, 2017; Wu & He, 2018). We set the shared distribution Ψ as a zero mean, unit variance Gaussian distribution $\mathcal{N}(0, I)$. We can normalize c_i to force each of the content embeddings c_i to follow $\mathcal{N}(0, I)$ as

$$\hat{c_i} = \frac{c_i - \mu(c_i)}{\sigma(c_i)} \tag{4}$$

where, $\mu(c_i)$ and $\sigma(c_i)$ are the mean and variance of c_i respectively. This operation shares the same formulation as Instance Normalization (IN). By combining IN and Equ. 1, we get the same formulation with AdaIN (Huang & Belongie, 2017).

As the generator usually have a multi-layer structure, we can use this C-S disentanglement design only in the first layer, which is denoted as first-layer framework. To increase the capacity of generator G, we can also inject it into the network at the front of each layer, which is denoted as multi-layer framework.

(a) (b) (c) (d) (e)

Figure 2: (a) Fused content and style. (b) first-layer. (c) first-layer w/o share content distribution. (d) unit-ball nomalization. (e) Instance normalization.

3.2 Demystifying C-S disentanglement

In this subsection, we perform some experiments to prove that assigning independent and different roles to content and style to model real data distribution is the key for C-S disentanglement.

If we fuse C-S first, and use the fused data to both model the distribution and predict parameters to modify the distribution, the network can hardly disentangle any meaningful information as show in Fig. 2 (a), even the multi-layer version is used. When we design the network following Fig. 1, as Fig. 2 (b) shows, our first-layer framework is able to disentangle the pose and identity. The content plays the role to model the shared distribution, when this contr is removed and leave the content unconstrained, the result is shown in Fig. 2 (c).

To better understand how the structure bias influences the disentanglement, we visualize the generated images and the information carried by the two paths during the training process in Fig. 3. Here we only show the images that have the same content embedding but different style embedding due to space limitation, more results can be found in the supplementary materials. We choose the KL divergence as the quantitative information capacity for the content and style path inspired by (Higgins et al., 2017). The information carried by the content path increases earlier than the style path. Meanwhile as the generated images show, the faces starts to rotate, which indicates pose is disentangled and extracted into the content path. Then the instance-specific features are extracted into the style embedding space.

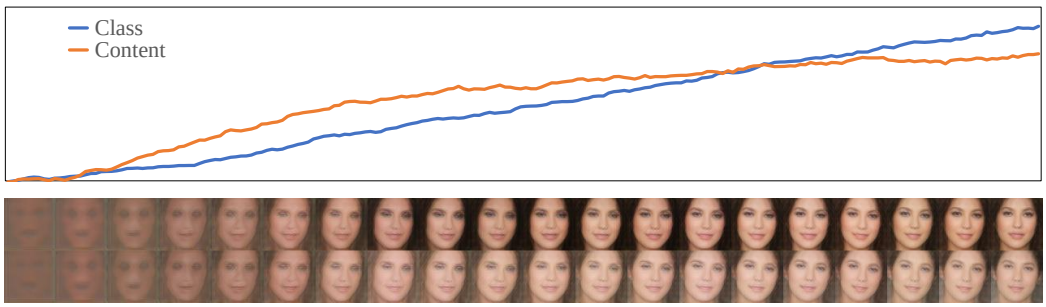


Figure 3: Visualization of the training process. Top: information carried by the content (blue) and style (orange) path. Bottom: generated images at different gradient update steps. The two images in each column have the same content embedding and different style embedding.

3.3 Techniques

Latent Optimization. In the style-content disentanglement literature, encoders are commonly used to extract the embeddings. While an encoder has a large parameter space, the model needs to optimize the encoder to adjust the latent embedding. During this optimization process, the inductive bias on the latent spaces is diluted by encoders. To maximum the bias, we adopt a latent optimization scheme (Bojanowski et al., 2018; Gabbay & Hoshen, 2020a) to directly update the latent spaces instead of an amortized fashion using the encoders. For the amortized scheme, there are leaks

(a) Fuses (b) Amortized (c) w/o R (d) Ours

Figure 4: Ablation study. R indicates reparametric module.

between style and content latent space, and the result is worse than latent optimization, as shown in Figure 4 (b) and (d).

Reparametric Module.

In our experiment, we find the reconstruction performance is poor, as shown in Figure 4 (c).

Inspired by VAE Kingma & Welling (2014), we design a reparametric module. Assume we have a mean embedding μ with a standard deviation σ, the reparametrized output is $\sigma X + \mu$, where $X \sim \mathcal{N}(0, 1)$. To further simplify the problem, we set $\sigma = 1$ following Wu et al. (2019c) and Gabbay & Hoshen (2020a). The reparametric module can make the latent space continuous which is helpful for backpropagation. Though we assume every image as is a distinct identity, the latent spaces are continuous after training, e.g. different style embeddings of the same person are close to each other, as shown in Figure 7. Note that the reparametric module must be applied both on the two paths of the style-based generator. Otherwise, the structure bias will fail to lead disentanglement on content and style. (see appendix).

Instance Discrimination. Instance discrimination can automatically discover apparent similarity among semantic categories (Wu et al., 2018). Inspired by this, we propose to use instance discrimination as a complementary constraint to guide the process of disentanglement. With instance discrimination, we first pretrain a ResNet-18 (He et al., 2016) Φ and define a collection of layers of Φ as $\{\Phi_l\}$. Among several representative methods (Wu et al., 2018; Ye et al., 2019; He et al., 2020), we observe that the method in Wu et al. (2018) achieves the best performance in our task. Given two images I_i and I_j, we mix the embeddings to generate $u = G(R(s_i), R(c_j))$ and $v = G(R(s_j), R(c_i))$. For samples sharing the same style embedding, we enforce the feature distance in Φ between them to be close. This loss term can be written as

$$\mathcal{L}_{\mathcal{ID}} = \sum_l \lambda_l(\|\Phi_l(u) - \Phi_l(x)\|_1 + \|\Phi_l(v) - \Phi_l(y)\|_1), \tag{5}$$

where $x = G(R(s_i), R(c_i))$ and $y = G(R(s_j), R(c_j))$. The hyperparameters $\{\lambda_l\}$ balance the contribution of each layer l to the loss.

3.4 Loss

Perceptual Loss. Perceptual Loss is widely used in weakly supervised and unsupervised methods in diverse areas (Wu et al., 2020; Ren et al., 2020; Wu et al., 2019c). Gabbay & Hoshen (2020a) claimed that perceptual loss is not an extra supervision for disentanglement. We adopt a VGG (Simonyan & Zisserman, 2015) perceptual loss $\mathcal{L}_{\mathcal{P}}$ as a reconstruction loss, implemented by Hoshen et al. (2019).

Information Bottleneck. Burgess et al. (2018a) propose improving the disentanglement in β-VAE by controlling the capacity increase, i,e., pressuring the KL divergence to be at a controllable value. This motivated us to control the information bottleneck capacity of different paths to enhance the model inductive bias. Similar to Anneal VAE (Burgess et al., 2018b), we introduce a information bottleneck given by

$$\mathcal{L}_{\mathcal{IB}} = \gamma_s\|s^2 - C_s\|_1 + \gamma_c\|c^2 - C_c\|_1 \tag{6}$$

where C_s and C_c are the information capacity controlling the amount of information of the style path and content path, respectively. During training, C_s and C_z increase linearly. The rate of increase

Table 1: Performance comparison in content tranfer metric. For Wu et al. (2019c), we provide pseduo ground truth landmarks.

Method	Supervision	Cars3D	Chairs	CelebA
DrNet (Denton & Birodkar, 2017)		0.146	0.294	0.221
Cycle-VAE (Jha et al., 2018)	✓	0.148	0.240	0.202
Lord (Gabbay & Hoshen, 2020a)		0.089	**0.121**	0.163
FactorVAE (Kim & Mnih, 2018)		0.190	0.287	0.369
Wu et al. (2019c)*	✗	–	–	0.185
Ours		**0.082**	0.190	**0.161**

is controlled by the increase steps and the maximum value. By controlling the increase rate, the content path is forced to encode information first, so that the learning process is more consistent with our assumptions about the data: the shared conditional variable y is learned first.

Full Objective. Our full objective is

$$w_P \mathcal{L}_P + w_{IB} \mathcal{L}_{IB} + w_{ID} \mathcal{L}_{ID}, \tag{7}$$

where hyperparameters w_P, w_{IB}, and w_{ID} represent the weights for each loss term.

4 Experiments

Dataset. We test our method on three datasets: **Car3D** (Reed et al., 2015), **Chairs** (Aubry et al., 2014), **CelebA** (Liu et al., 2015). Car3D contains 183 car models, each rendered from fixed 96 poses. Chairs consist of 1393 chair models, each rendered from fixed 62 poses. CelebA contains 202,599 facial images of 10,177 celebrities. For more results (including other datasets), please refer to Appendix C.

Baselines. Among all the prior works, we choose several SOTA class-supervised content and style disentanglement benchmarks for comparisons: Cycle-VAE (Jha et al., 2018), a variant of VAE using cycle-consistency; DrNet (Denton & Birodkar, 2017), an adversarial approach; Lord (Gabbay & Hoshen, 2020a), a latent optimization method. We also choose two unsupervised disentanglement techniques: FactorVAE (Kim & Mnih, 2018), a method encourages the distribution of representations to be factorial; Wu et al. (2019c) [1], a two-branch VAE framework based on unsupervised structure learning. More details for baselines will be presented in Appendix B.

4.1 Quantitative Experiments

We compare our method with the baselines on Car3D, Chairs and CelebA.

Content Transfer Metric. To demonstrate the ability of disentanglement of our method, we follow the protocol of Gabbay & Hoshen (2020a) to measure the quality of content transfer by LPIPS (Zhang et al., 2018). For Car3D and Chairs datasets, the content labels (ground truth) are available. For images x and y sampled from testing set randomly, we compute LPIPS between $G(C_x, S_y)$ and the corresponding ground truth from the same class of y. For CelebA, we randomly sample two images x and y of the same identity, and we retrieve image that has the most similar pose z from testse, i.e. the nearest neighbour in the 68 facial-landmarks space. We measure the similarity between y and $G(C_x, S_z)$. As shown in Table 1, our method outperform other methods on Car3D and CelebA datasets. We also achieve good performance on Chairs dataset. Note that we are fully unsupervised, the methods with supervision have dense poses for the same identity.

Classification Metric. Classification accuracy is used to evaluate disentanglement in several literature (Denton & Birodkar, 2017; Jha et al., 2018; Gabbay & Hoshen, 2020a). Following Jha et al.

[1] There is no open-sourced implementation for it. We modify `https://github.com/CompVis/vunet` and provide pseudo ground truth landmarks to the network.

Table 2: Performance comparison in classification metric. For Wu et al. (2019c), we provide pseduo ground truth landmarks.

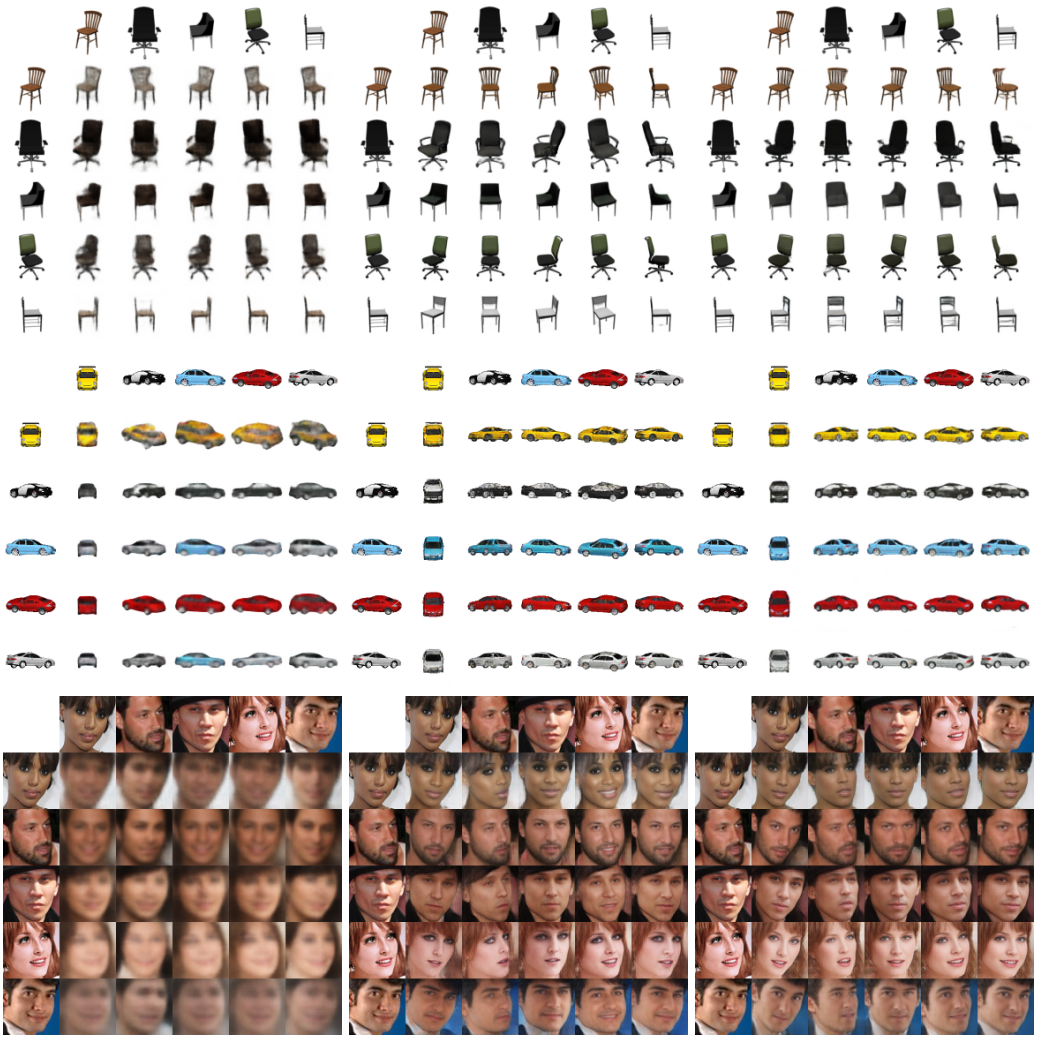
Method	Supervision	Cars3D		Chairs		CelebA	
		$y \rightarrow c$	$y \leftarrow c$	$y \rightarrow c$	$y \leftarrow c$	$y \rightarrow c$	$y \leftarrow c$
DrNet (Denton & Birodkar, 2017)		0.27	**0.03**	0.06	0.01	4.99	**0.00**
Cycle-VAE (Jha et al., 2018)	✓	0.81	0.77	0.60	0.01	2.80	0.12
Lord (Gabbay & Hoshen, 2020a)		**0.03**	0.09	**0.02**	0.01	4.42	0.01
FactorVAE (Kim & Mnih, 2018)		0.07	0.01	0.14	0.01	5.34	**0.00**
Wu et al. (2019c)*	✗	–	–	–	–	**5.42**	0.11
Ours		0.33	0.24	0.66	0.05	4.15	0.05

Figure 5: Linear interpolation results between the top-left and the bottom-right corner. Content is constant in the rows, while style is constant in the columns.

(2018), we train two single fully-connected layer models to classify content label from style latent space and classify identity label from content latent space. All the two subspace is composed of the whole dataset. Due to no content label of CelebA, we regress the position of 68 facial landmarks using a fully-connected layer model. The results are summarized in Table 2. Though without supervision, our method is comparable to several methods. We observe that the classification metric is influenced by three factors: 1) Information capacity. For some methods, e.g. FactorVAE (Kim & Mnih, 2018), the poor reconstruction quality leads to small amount of information encoded. And thus the classification accuracy naturally is low. 2) Dimension. The dimensions of the latent vectors of different methods vary from ten to hundreds. Actually, high dimension usually leads to high accuracy for containing more information. 3) Leak. Finally, this metric is related to disentanglement ability. However, influenced by the other two factors, we doubt if classification metric could give a fair measure for disentanglement.

4.2 Qualitative Experiments

Interpolation. In Figure 5, we conduct linear interpolation to capture the variation in the two latent manifolds. From left to right, the identity is changed smoothly with interpolated style latent space while maintaining the pose information. Analogously, the identity remains the same while changing the pose from top to down. As shown in Figure 7, one interesting observation is that zero content codes indicate canonical view without any specification. We also show zero style codes that are mean shape learned from the dataset.



FactorVAE Lord Ours

Figure 6: Visual analogy results comparing our method and baselines on Chairs, Car3D and CelebA (from top to bottom).

Retrieval. To demonstrate the latent space learned by our method, we demonstrate the result by performing nearest neighbor retrieval experiments, as shown in Figure 7.

Visual Analogy & Comparison. Visual analogy (Reed et al., 2015) is to switching between style (left column) and content (top row) codes for each pair within a set of test images. We show the visual analogy result of our method against FactorVAE (Kim & Mnih, 2018) (typical unsupervised baseline) and Lord (strongest baseline with group observation) in Figure 6 on Chairs, Car3D and CelebA. The results of FactorVAE on all dataset have poor reconstruction and wrong content transfer. On Cars3D, Lord has artifacts (e.g. third column) and could not capture the real style of the test images (e.g. forth row). On CelebA, the transfer result of Lord has ambiguity. E.g for the fifth column, all the images follow the same expression. While for the fifth row, the images do not follow the expression from the content code. Our method achieves excellent pose transfer and maintain the identity of a single image.

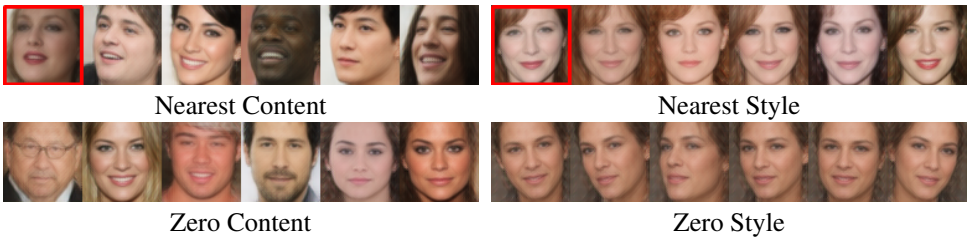


Figure 7: More exploration for content and style latent space. The red box is the key image to perform retrieval. When we visualize nearest style embedding retrieval, we keep content embedding as zero for better demonstration.

Figure 8: Results on Market-1501 dataset.

5 Extension

Representation disentanglement for unsupervised domain translation are well exploited in literature (Gabbay & Hoshen, 2020a; Wu et al., 2019b; Gabbay & Hoshen, 2020b). In this work, we explore an additional application for style-content disentanglement. For 3D reconstruction, signle-view settings lack reliable 3D constraints which can cause unresolvable ambiguities (Wu et al., 2019a). We conduct extensive experiments to show our method helps 3D reconstruction task.

On Chairs, we adopt Pix2Vox (Xie et al., 2019), a framework for single-view and multi-view 3D reconstruction to recover the shape of 3D object. In this setting, we have the chance to show the advantages of multi-view over single-view, which shows that our model retains the identity information in different views. As shown in Figure 9, the 3D objects generated from multi-view inputs based on our method are much better than those generated from single view. And our results are comparable to those generated from ground-truth multi-view images. For results on Celeba, please refer to Appendix D.

6 Conclusion

We present an unsupervised content and style disentanglement method. Our method can automatically disentangle the space into two subspaces with our formulation on the latent space, combining latent optimization. Our method outperforms unsupervised approaches and is comparable to or even better than the state-of-the-art supervised methods. Additionally, we use our method to help single-view 3D reconstruction to show the excellent disentanglement ability of our method. For the limitation, we fail to work on datasets with large appearance variance like CIFAR-10. As for future work, our method could be adopted to more downstream tasks, e.g., domain translation. An interesting direction is to inject our method into the instance discrimination pipeline. With disentangled representation, contrastive learning can be performed more properly.

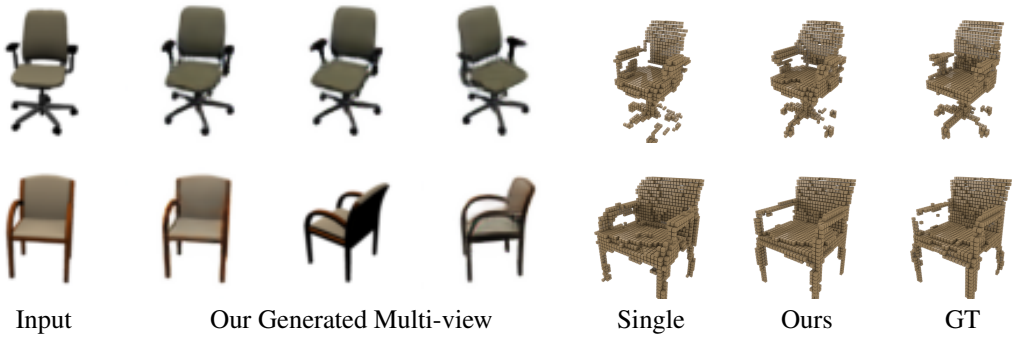


Figure 9: 3D reconstructions on Chairs. Input is augmented to multi-view inputs by our method. Single represents the object generated by only Input. Ours represents the object generated by multi-view inputs. GT represents the object generated by ground truth of multi-view inputs.

REFERENCES

Yazeed Alharbi and Peter Wonka. Disentangled image generation through structured noise injection. In *CVPR*, 2020.

Mathieu Aubry, Daniel Maturana, Alexei A. Efros, Bryan C. Russell, and Josef Sivic. Seeing 3d chairs: Exemplar part-based 2d-3d alignment using a large dataset of CAD models. In *CVPR*, 2014.

Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam. Optimizing the latent space of generative networks. In *ICML*, 2018.

Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *AAAI*, 2018.

Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *ICCV*, 2017.

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

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in beta-vae. *arXiv preprint arXiv:1804.03599*, 2018b.

Brian Chao. A collection of high-quality anime faces. `https://github.com/bchao1/Anime-Face-Dataset`, 2019.

Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *NeurPIS*, 2018.

Emily L. Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. In *NeurIPS*, 2017.

Patrick Esser, Robin Rombach, and Björn Ommer. A disentangling invertible interpretation network for explaining latent representations. In *CVPR*, 2020.

Aviv Gabbay and Yedid Hoshen. Demystifying inter-class disentanglement. In *ICLR*, 2020a.

Aviv Gabbay and Yedid Hoshen. Improving style-content disentanglement in image-to-image translation. *CoRR*, abs/2007.04964, 2020b.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

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 *ICLR*, 2017.

Yedid Hoshen, Ke Li, and Jitendra Malik. Non-adversarial image synthesis with generative latent nearest neighbors. In *CVPR*, 2019.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

Ananya Harsh Jha, Saket Anand, Maneesh Singh, and V. S. R. Veeravasarapu. Disentangling factors of variation with cycle-consistent variational auto-encoders. In *ECCV*, 2018.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

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

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

Dmytro Kotovenko, Artsiom Sanakoyeu, Sabine Lang, and Bjorn Ommer. Content and style disentanglement for artistic style transfer. In *ICCV*, 2019.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*, 2017.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Chuan Li and Michael Wand. Combining markov random fields and convolutional neural networks for image synthesis. In *CVPR*, 2016.

Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. In *IJCAI*, 2017.

Yu-Jhe Li, Ci-Siang Lin, Yan-Bo Lin, and Yu-Chiang Frank Wang. Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In *ICCV*, 2019.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *ICML*, 2019.

Michaël Mathieu, Junbo Jake Zhao, Pablo Sprechmann, Aditya Ramesh, and Yann LeCun. Disentangling factors of variation in deep representation using adversarial training. In *NeurIPS*, 2016.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, 2012.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Stanislav Pidhorskyi, Donald A. Adjeroh, and Gianfranco Doretto. Adversarial latent autoencoders. In *CVPR*, 2020.

Scott E. Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. In *NeurIPS*, 2015.

Xuanchi Ren, Haoran Li, Zijian Huang, and Qifeng Chen. Self-supervised dance video synthesis conditioned on music. In *ACM MM*, 2020.

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, 2020.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

Attila Szabó, Qiyang Hu, Tiziano Portenier, Matthias Zwicker, and Paolo Favaro. Challenges in disentangling independent factors of variation. In *ICLRW*, 2018.

Fanzi Wu, Linchao Bao, Yajing Chen, Yonggen Ling, Yibing Song, Songnan Li, King Ngi Ngan, and Wei Liu. Mvf-net: Multi-view 3d face morphable model regression. In *CVPR*, 2019a.

Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *CVPR*, 2020.

Wayne Wu, Kaidi Cao, Cheng Li, Chen Qian, and Chen Change Loy. Transgaga: Geometry-aware unsupervised image-to-image translation. In *CVPR*, 2019b.

Wayne Wu, Kaidi Cao, Cheng Li, Chen Qian, and Chen Change Loy. Disentangling content and style via unsupervised geometry distillation. In *ICLRW*, 2019c.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018.

Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *ICCV*, 2019.

Xianglei Xing, Tian Han, Ruiqi Gao, Song-Chun Zhu, and Ying Nian Wu. Unsupervised disentangling of appearance and geometry by deformable generator network. In *CVPR*, 2019.

Mang Ye, Xu Zhang, Pong C. Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *CVPR*, 2019.

Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *CVPR*, 2014.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Weiwei Zhang, Jian Sun, and Xiaoou Tang. Cat head detection - how to effectively exploit shape and texture features. In *ECCV*, 2008.

Xingyi Zhou, Arjun Karpur, Linjie Luo, and Qixing Huang. Starmap for category-agnostic keypoint and viewpoint estimation. 2018.

Yang Zou, Xiaodong Yang, Zhiding Yu, BVK Kumar, and Jan Kautz. Joint disentangling and adaptation for cross-domain person re-identification. In *ECCV*, 2020.

A Implementation Details

For the style-based generator, we use an implementation from Gabbay & Hoshen (2020a). All the models are trained on an Nvidia Tesla V100 GPU. The model is implemented in PyTorch (Paszke et al., 2017). For Car3D, Chairs and Celeba, we set the size of the style embedding d_s to 256, the size of the content embedding d_c to 128 and the epoch to 200. All the images are resized to 64×64. For the hyperparameters, $\{\lambda_l\}$ are set to be $[1, 1, 1, 1, 1]$ and $w_P = 5, w_{IB} = 1, w_{ID} = 1$. We used Adam (Kingma & Ba, 2015) with a learning rate of 0.0003 for the models and 0.003 for the latent spaces.

For the information bottleneck, by taking the training process of the model without the information bottleneck as a reference, we determine the increase steps and the maximum of the information capacity Cc and Cs. We can enhance the model inductive bias by tuning these parameters. For Chairs, we set the maximum of Cc to 5, the start value of Cc to 2 and the increase steps of Cc to 1.4×10^5. Note that our model achieve the state-of-the-art performance on Chairs even without information bottleneck.

B BASELINE DETAILS

We use open-sourced implementations for Cycle-VAE (Jha et al., 2018) [2], DrNet (Denton & Birodkar, 2017) [3], Lord (Gabbay & Hoshen, 2020a) [4] and FactorVAE (Kim & Mnih, 2018) [5].

For FactorVAE, we traverse the latent space to select the dimensions related to pose as content embedding and treat the other dimensions as style embedding. For Wu et al. (2019c), there is no open-sourced implementation. We use the code from `https://github.com/CompVis/vunet` which use ground truth landmarks as input instead of learning the landmarks unsupervisedly. To achieve the pseudo ground truth landmarks, we use face detection library (Bulat & Tzimiropoulos, 2017) for Celeba and object keypoints detection library (Zhou et al., 2018) for Car3D and Chairs. We try to use L1 and perceptual loss for all the baselines and select the best.

We split the datasets into training and testing sets. For Celeba, we randomly select 1000 among 10177 celebrities for testing. For Car3D, we randomly select 20 among 183 CAD models for testing. For Chairs, we randomly select 100 among 1393 models for testing. For baselines with group supervision, only the training sets are used for training. For unsupervised baselines and our method, all the datasets are used for training.

C MORE RESULTS

In this section, we demonstrate more qualitative comparison and more qualitative results (including more datasets).

In the main paper, for unsupervised baselines, we only compare our method with FactorVAE (Kim & Mnih, 2018) limited to space. As shown in Figure 10, we also outperform Wu et al. (2019c). For Wu et al. (2019c), the disentanglement is poor, such that the content embeddings control almost all the factors while the style embeddings control the tone.

For datsets in the main paper, We provide more qualitative results in Fig. 17, 18, 19, 20 and 21. Moreover, we also apply our method on higher resolution images and achieve good performance, as shown in Figure 13.

Besides the datasets introduced in the main paper, we make additional experiments on other datasets: such as **MNIST** (LeCun et al., 2010), **Cat** (Parkhi et al., 2012; Zhang et al., 2008), **Anime** (Chao, 2019) and **Edges2Shoes** (Yu & Grauman, 2014). MNIST has 70k examples for 10 handwritten digits. Cat has 1.2k cat head images. Anime contains 63,632 anime faces. Finally, Edges2Shoes is a dataset consisting of 50k paired shoe and edge map images. The results are shown in Figure 14, 15 ,16 and 14.

D MORE 3D RECONSTRUCTION

Our setting treat every image as a single identity (style) without bring ambiguity for augment single-view images. On Celeba, We use MVF-Net (Wu et al., 2019a) based on multi-view to reconstruct 3D facial shapes. For a given image, we can get the corresponding style embedding content embedding. Then we can get front, left and right view of this image combining the extracted style embedding and prepared content embeddings [6]. As shown in Figure 11, our augmented multi-view images are consistent and the 3D meshes based on our method are more accurate than those based on Lord.

E MORE ABLATION STUDY

In the main paper, we do not perform enough ablation study for the proposed modules.

[2] `https://github.com/ananyahjha93/cycle-consistent-vae`

[3] `https://github.com/ap229997/DRNET`

[4] `https://github.com/avivga/lord-pytorch`

[5] `https://github.com/1Konny/FactorVAE`

[6] We retrieve the nearest neighborhood (facial landmarks space) of suggested inputs of MVF-Net and extract content embedding.

Wu et al. Our

Figure 10: Comparison between Wu et al. and our method. For Wu et al., the images are mainly determined by content embeddings, while style embeddings only change the tone.

Ours ⟵ Input ⟶ Lord

Figure 11: 3D face reconstruction. Given an image, we first generate multi-view images and then use them as augmented input.

For instance discriminator, we

F Structure Bias vs. Group Observation

G Unseen Images Inference

Though we learn to disentangle in an unsupervised manner, we may need to process unseen images at test time. An natural solution is to train encoders to encode images to the latent spaces. We train style encoder E_s and content encoder E_c by

$$\mathcal{L}_E = \|E_s(I_i) - s_i\|_1 + \|E_c(I_i) - c_i\|_1. \tag{8}$$

After training, we applied our model trained on the CelebA dataset to faces collected by Wu et al. (2020) including paintings and cartoon drawings. As shown in Figure 12, our method still works well even though it has never seen such images during training.

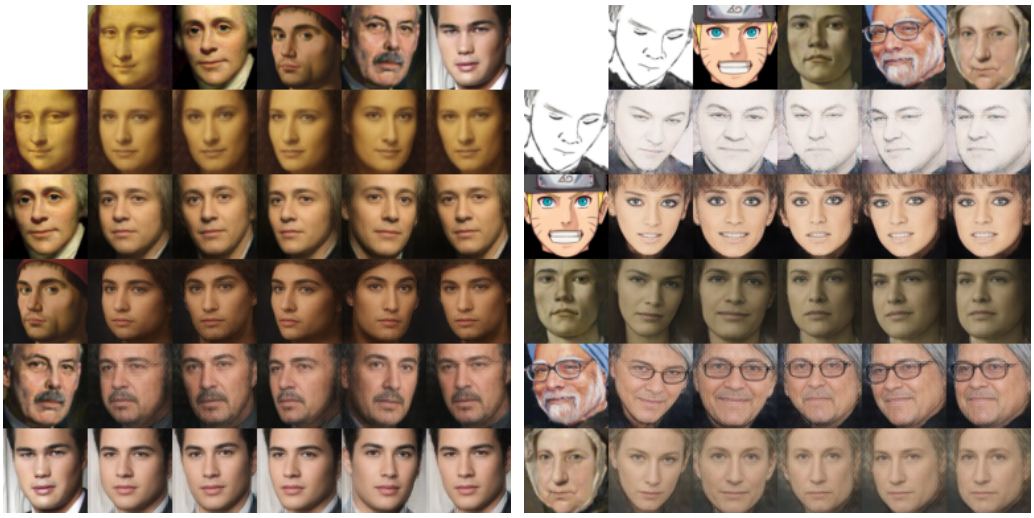


Figure 12: Inference for unseen images. Our method performs well on images from different domains: painting and cartoon.

Table 3: Ablation study for infomation bottleneck on Chairs dataset.

Method	Content transfer	Classification	
		$y \leftarrow c$	$y \rightarrow c$
Ours (w/o Information Bottleneck)	0.148	0.33	0.24
Ours	0.148	0.33	0.24

Table 4: Ablation study for instance discrimination on Celeba dataset.

Method	Content transfer	Classification	
		$y \leftarrow c$	$y \rightarrow c$
Ours (w/o Instance Discrimination)	0.148	0.33	0.24
Ours	0.148	0.33	0.24

Figure 13: Results on Celeba with 128×128 resolution. Zoom in for better view.

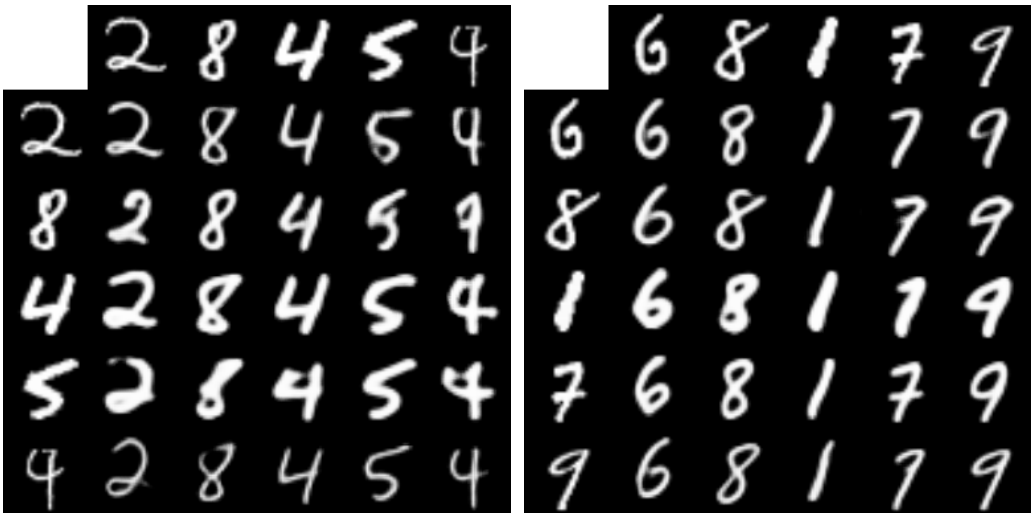


Figure 14: Results on MNIST dataset. Content indicates geometric attribute and style indicates texture.

Figure 15: Results on Cat dataset. Content indicates pose and style indicates identity. The result further qualitatively demonstrate the ability of our method to disentangle on real-world data.

Figure 16: Results on Anima dataset. Content indicates pose and style indicates identity.

Figure 17: More visual anology of our method on Car3D.

Figure 18: More visual anology of our method on Car3D.

Figure 19: More visual anology of our method on Chairs.

Figure 20: More visual anology of our method on Chairs.

Figure 21: More visual anology of our method on Celeba.