

Rethinking Content and Style: Exploring Bias for Unsupervised Disentanglement

Xuanchi Ren^{1*} Tao Yang^{2,3} Yuwang Wang^{3†} Wenjun Zeng³
¹HKUST ²Xi'an Jiaotong University ³Microsoft Research Asia

Abstract

Content and style (C-S) disentanglement intends to decompose the underlying explanatory factors of objects into two independent subspaces. From the unsupervised disentanglement perspective, we rethink content and style and propose a formulation for unsupervised C-S disentanglement based on our assumption that different factors are of different importance and popularity for image reconstruction, which serves as a data bias. The corresponding model inductive bias is introduced by our proposed C-S disentanglement Module (C-S DisMo), which assigns different and independent roles to content and style when approximating the real data distributions. Specifically, each content embedding from the dataset, which encodes the most dominant factors for image reconstruction, is assumed to be sampled from a shared distribution across the dataset. The style embedding for a particular image, encoding the remaining factors, is used to customize the shared distribution through an affine transformation. The experiments on several popular datasets demonstrate that our method achieves the state-of-the-art unsupervised C-S disentanglement, which is comparable or even better than supervised methods. We verify the effectiveness of our method by downstream tasks: image-to-image translation and single-view 3D reconstruction. Project page at <https://github.com/xrenaa/CS-DisMo>.

1. Introduction

The disentanglement task aims to recover the underlying explanatory factors of natural images into different dimensions of latent space, and provide an informative representation for downstream tasks like image translation [35, 17], domain adaptation [21] and geometric attributes extraction [38], etc.

In this work, we focus on content and style (C-S) disentanglement, where content and style represent two independent latent subspaces. Most of the previous C-S disentanglement

works [5, 12, 2, 6] rely on supervision. For example, Gabbay and Hoshen [6] achieve disentanglement by forcing images from the *same group* to share a common embedding. It is not tractable, however, to collect such a dataset (*e.g.* groups of paintings with each group depicting the same scene in different styles, or groups of portraits with each group depicting the same person with different poses). To our best knowledge, the only exception is Wu et al. [34] which, however, forces the content to encode pre-defined geometric structure limited by the expressive ability of 2D landmarks.

Previous works define the content and style based on either the supervision or manually pre-defined attributes. There is no general definition of content and style for *unsupervised* C-S disentanglement. In this work, we define content and style from the perspective of VAE-based unsupervised disentanglement works [9, 3, 14, 4]. These methods try to explain the images with the latent factors, of which each controls only one interpretable aspect of the images. However, extracting all disentangled factors is a very challenging task, and Locatello et al. [25] prove that unsupervised disentanglement is fundamentally impossible without inductive bias on both the model and data. Furthermore, these methods have limited down-stream applications due to poor image generation quality on real-world datasets.

Inspired by the observation that the latent factors have different degrees of importance for image reconstruction [3], we assume the disentangled factors are of different importance when modeling the real data distributions. Instead of finding all the independent factors, which is challenging, we make the problem tractable by defining *content* as a group of factors that are the most important ones for image reconstruction across the whole dataset, and defining *style* as the remaining ones. Take the human face dataset CelebA [24] as an example, as pose is a more dominant factor than identity for image reconstruction across the face dataset, content encodes pose, and identity is encoded by style. We further assume that each content embedding of the dataset is sampled from a shared distribution, which characterizes the intrinsic property of content. For the real-world dataset CelebA, the shared distribution of content (pose) is approximately a Standard Normal Distribution, where zero-valued embed-

*This work was done when Xuanchi Ren was an intern at MSRA

†Corresponding author

ding stands for the canonical pose. For the synthetic dataset Chairs [1], as each image is synthesized from equally distributed surround views, the shared distribution of content (pose) is approximately an Uniform Distribution.

Based on the above definitions and assumptions, we propose a problem formulation for unsupervised C-S disentanglement, and a C-S **DisMo** (C-S Disentanglement Module) which assigns different and independent roles to content and style when approximating the real data distributions. Specifically, C-S DisMo forces the content embeddings of individual images to follow a common distribution, and the style embeddings are used to scale and shift the common distribution to match the target image distribution via a generator. With the above assumptions as the data inductive bias, and C-S DisMo as the corresponding model inductive bias, we achieve unsupervised C-S disentanglement with good image generation quality. Furthermore, we demonstrate the effectiveness of our disentangled C-S representations on two down-stream applications, *i.e.*, image-to-image translation and single-view 3D reconstruction.

We follow Gabbay and Hoshen [6] to apply latent optimization to optimize the embeddings and the parameters of the generator. Please note that we only use the image reconstruction loss as the supervision; no human annotation is needed. We also propose to use instance discrimination as an auxiliary constraint to assist the disentanglement.

The experiments on several popular datasets demonstrate that our method achieves the state-of-the-art (SOTA) unsupervised C-S disentanglement, which is comparable or even better than supervised methods. Furthermore, by simplifying the factors disentanglement problem into the C-S disentanglement problem, we achieve much better performance than the SOTA VAE-based unsupervised disentanglement method when modified for C-S disentanglement by manually splitting the factors into content and style.

Main contributions. The main contributions of our work are as follows: (i) By rethinking content and style from the perspective of VAE-based unsupervised disentanglement, we achieve unsupervised C-S disentanglement by introducing both data and model inductive bias. (ii) We propose the C-S DisMo to assign different and independent roles to content and style when modeling the real data distributions, and we provide several solutions for the distribution constraint of content. (iii) We verify the effectiveness of our method by presenting two down-stream applications based on the well-disentangled content and style.

2. Related Work

Unsupervised Disentanglement. There have been a lot of studies on unsupervised disentangled representation learning [9, 3, 14, 4]. These models learn disentangled factors by factorizing aggregated posterior. However, Locatello et al. [25] prove that unsupervised disentanglement is impos-

sible without introducing inductive bias on both the models and data. Therefore, these models are currently unable to obtain a promising disentangled representation. Inspired by these previous unsupervised disentanglement works, we revisit and formulate the unsupervised C-S disentanglement problem. We simplify the challenging task of extracting individual disentanglement factors into a manageable task: disentangling two groups of factors (content and style).

C-S Disentanglement. Originated from style transfer, most of the prior works on C-S disentanglement divide latent variables into two spaces relying on group supervision. To achieve disentanglement, Mathieu et al. [26] and Szabo et al. [31] combine the adversarial constraint and auto-encoders. Meanwhile, VAE [16] is combined with non-adversarial constraints, such as cycle consistency [12] and evidence accumulation [2]. Furthermore, latent optimization is shown to be superior to amortized inference for C-S disentanglement [6]. The only exception is Wu et al. [34], which proposes a variational U-Net with structure learning for disentanglement in an unsupervised manner, but is limited by the expressive ability of 2D landmarks. In our work, we focus on the unsupervised C-S disentanglement problem and explore inductive bias for unsupervised disentanglement.

Image Translation. Image translation [11, 22] between domains tries to decompose the latent space into domain-shared representations and domain-specific representations with the domain label of each image as supervision. The decomposition relies on the “swapping” operation and pixel-level adversarial loss without semantic level disentanglement ability. This pipeline fails in the unsupervised C-S disentanglement task on the single domain dataset due to lack of domain supervision. Our unsupervised C-S disentanglement task is to disentangle the latent space into content (containing most dominant factors typically carrying high-level semantic attributes) and style (containing the rest of the factors). We achieve disentangled content and style by assigning different roles to them without relying on domain supervision or the “swapping” operation. We formulate the problem for a single domain but it can be extended to cross-domain to achieve domain translation without domain supervision, as shown in Figure 10.

3. Exploring Inductive Bias for Unsupervised C-S Disentanglement

3.1. Problem Formulation

For a given dataset $\mathcal{D} = \{I_i\}_{i=1}^N$, where N is the total number of images, we assume each image I_i is sampled from a distribution $P(x|\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_k)$, where $\{\mathbf{f}_i\}_{i=1}^k$ are the disentangled factors. Disentangling all these factors unsupervisedly is a challenging task, which has been proved to be fundamentally impossible without introducing the model and data inductive bias [25]. Based on the observation that

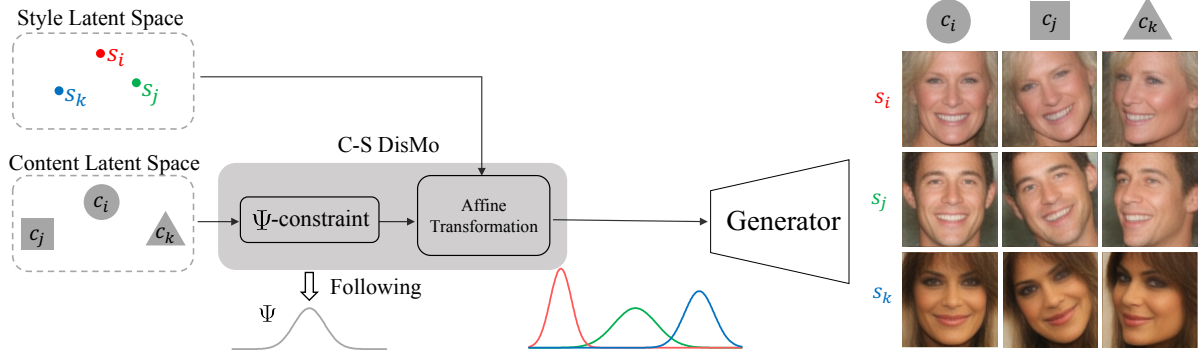


Figure 1. Overview of our method. Content embeddings c_i, c_j, c_k are labelled with different shapes, and style embeddings s_i, s_j, s_k are labelled with different colors. A C-S Disentanglement Module (C-S DisMo) is composed of a Ψ -constraint and an affine transformation. The Ψ -constraint forces content embeddings to follow a shared distribution Ψ and the affine transformation scales and shifts the shared content distribution with different styles (colors) as the Generator’s input to approximate the target image distributions. Each image from 3×3 grids (right side) is generated with the content embedding from the column and style embedding from the row.

the factors play roles of different importance for image reconstruction [3], we assume $\{f_i\}_{i=1}^k$ are of different importance and popularity for modeling the image distribution P . We define the *content* c as representing the most important factors across the whole dataset for image reconstruction and *style* s as representing the rest ones. We assume c follows a shared distribution across the whole dataset, and assign each image I_i a style embedding s_i which parameterizes P to be an image-specific distribution $P_{s_i}(x|c)$. This serves as the data bias for our unsupervised C-S disentanglement.

With a generator G_θ that maps content and style embeddings to images, where θ is the parameter of the generator, we further parameterize the target image distributions as $\{\hat{P}_{\theta, s_i}(x|c)\}_{i=1}^N$. For each image I_i , we assign c_i as the content embedding. All the content embeddings $\{c_i\}_{i=1}^N$ should conform the assumed distribution of content c , which is denoted as Ψ . Then we are able to estimate the likelihood of I_i by $\hat{P}_{\theta, s_i}(x|c = c_i)$. Given the dataset D , our goal is to minimize the negative log-likelihood of \hat{P} :

$$\min_{\theta, c_i, s_i} - \sum_{i=1}^N \log \hat{P}_{\theta, s_i}(x = I_i | c = c_i) - \sum_{i=1}^N \log \Psi(c = c_i). \quad (1)$$

3.2. Proposed C-S Disentanglement Module

Here we propose a framework to address the formulated problem in Section 3.1. We design a C-S Disentanglement Module (C-S DisMo) to assign different roles to content and style in modeling real data distributions according to their definitions (data bias) in Section 3.1, which serves as the corresponding model bias.

More specifically, as shown in Figure 1, a C-S DisMo is composed of a Ψ -constraint to enforce content embeddings to conform to Ψ , which corresponds to the second term in Eq. 1, and an affine transformation serving to customize the shared content distribution into image-specific distributions.

This module is followed by the generator to generate the target image.

The *affine transformation* is inspired by the observation that the mean and variance of features carry individual information [7, 19, 20, 10]. We use the style embeddings to provide the statistics to scale and shift content embeddings as

$$z_i = f_\sigma(s_i) \cdot c_i + f_\mu(s_i), \quad (2)$$

where f_σ and f_μ are two fully connected layers predicting the scalars for scaling and shifting respectively. When $\hat{P}_{\theta, s_i}(x|c = c_i)$ is a Normal Distribution, Eq. 1 is equivalent to minimizing:

$$\mathcal{L}_{CS} = \sum_{i=1}^N \|I_i - G_\theta(z_i)\| - \sum_{i=1}^N \log \Psi(c = c_i), \quad (3)$$

with the proof provided in the supplementary material.

For the reconstruction term in Eq. 3, we adopt a VGG perceptual loss [30, 29], which is widely used in unsupervised disentanglement methods [33, 34].

For the Ψ -constraint, i.e. the second term in Eq. 3, we propose and study discrimination-based, NLL-based and normalization-based solutions. Choosing the form of Ψ , which can better approximate the ground truth content distribution of the dataset, can result in better disentanglement. For real-world datasets, content is affected by a large number of random variables, we could assume the distribution of the content is nearly standard Normal Distribution. We describe details of these solutions and related limitations according to the form of Ψ below.

Discrimination-based solution can be adopted when Ψ has a tractable form for sampling. Inspired by adversarial learning [13], we propose to use a discriminator to distinguish between content embeddings $\{c_i\}_{i=1}^N$ (false samples) and items $\{\hat{c}_i\}_{i=1}^N$ sampled from Ψ (true samples). When it is difficult for the discriminator to distinguish true from false, the content embeddings are likely to follow Ψ .

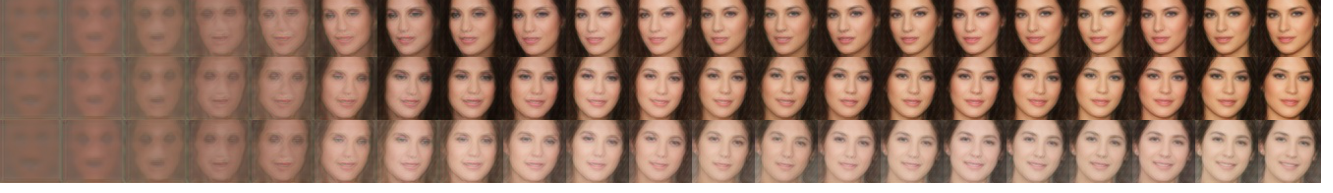


Figure 2. Generated images at different training steps. The first and second rows share the same style embedding. The second and third rows share the same content embedding.

NLL-based solution is inspired by flow-based generative models [15], and can be adopted when $\Psi = \mathcal{N}(\mu, \sigma^2)$. We can use negative log-likelihood (NLL) to optimize $\{c_i\}_{i=1}^N$ to follow Ψ as

$$\min_{c_i} \frac{1}{N} \sum_{i=1}^N \left(-\frac{\log 2\pi}{2} - \log \sigma - \frac{(c_i - \mu)^2}{2 \exp(2 \log \sigma)} \right). \quad (4)$$

Normalization-based solution can be adopted when Ψ has one of the following specific forms: *i*) a Standard Normal Distribution $\mathcal{N}(0, I)$, and *ii*) a Uniform Distribution. To approximately follow the $\mathcal{N}(0, I)$ constraint, Instance Normalization (IN) is used to force the mean and variance of c_i to be zeros and I respectively. When Ψ is a Uniform Distribution, we can use \mathcal{L}_2 normalization to force $\{c_i\}_{i=1}^N$ to follow Uniform Distribution approximately [27].

For these solutions, we show the qualitative and quantitative comparisons in Figure 3 and Table 3 respectively to verify their effectiveness. Furthermore, discrimination-based and negative log-likelihood (NLL)-based solutions need extra optimization terms which introduce instability. In our work, we mainly adopt normalization-based solution to meet the Ψ -constraint.

As shown in Figure 1, we can use the C-S DisMo before the generator, denoted as the Single C-S DisMo framework. We can also insert it before each layer of the generator to provide multiple paths for disentanglement, denoted as the Multiple C-S DisMo framework.

3.3. Demystifying C-S Disentanglement

In this section, we perform some experiments to verify that the C-S disentanglement is achieved by introducing inductive bias on model (C-S DisMo) and data (our assumptions of the dataset). The experimental setting can be found in Section 4.

To understand how C-S DisMo achieves disentanglement, we visualize the generated images during the training process of CelebA in Figure 2. As the generated images show, a mean shape of faces is first learned. Then the faces start to rotate, which indicates the pose, as a dominant factor for generation, is disentangled as content. After that, the identity features emerge since the identity is disentangled as style for better image generation.

If we treat content and style equally, *i.e.*, concatenating content and style embedding as the input of the generator,

the network can hardly disentangle any meaningful information for the CelebA dataset, as shown in Figure 3 (a). Our Single C-S DisMo framework with different solutions to meet Ψ -constraint can disentangle the content (pose) and style (identity) of human faces, as shown in Figure 3 (c)-(e). When Ψ -constraint is removed from C-S DisMo, the result is shown in Figure 3 (b), where the pose and identity can not be disentangled. For the Multiple C-S DisMo framework, as multiple paths are provided, and the network has more flexibility to approximate the target image distribution, it outperforms the Single C-S DisMo framework, as shown in Figure 3 (f).

We conduct experiments to demonstrate that better disentanglement can be achieved by choosing a better form for Ψ . For the real-world dataset CelebA, the distribution of pose is better modeled as a Standard Normal Distribution. As Figure 4 (a) and (b) show, IN achieves better disentanglement than \mathcal{L}_2 . For the synthetic Chairs [1] dataset, the distribution of pose is close to Uniform Distribution rather than Standard Normal Distribution. Figure 4 (c) and (d) show that the \mathcal{L}_2 normalization results in better consistency of identity and pose.

3.4. Auxiliary Loss Function

In addition to the \mathcal{L}_{CS} in Eq. 3, we propose two auxiliary loss functions to help the model to better disentangle C-S.

Instance discrimination. Instance discrimination can discover image-specific features [36]. The image-specific feature corresponds to style according to our definition. Inspired by this, we first pretrain a ResNet-18 [8] Φ unsupervisedly with the method in [36] and define a collection of layers of Φ as $\{\Phi_l\}$. Given two images I_i and I_j , we mix the embeddings to generate $u = G_\theta(s_i, c_j)$ and $v = G_\theta(s_j, c_i)$. 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

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

where $x = G_\theta(s_i, c_i)$ and $y = G_\theta(s_j, c_j)$. The hyperparameters $\{\lambda_l\}$ balance the contribution of each layer l to the loss. $\{\lambda_l\}$ are set to be $[1, 1, 1, 1, 1]$.

Information bottleneck. Burgess et al. [3] propose improving the disentanglement by controlling the capacity increment. This motivated us to control the information bottle-

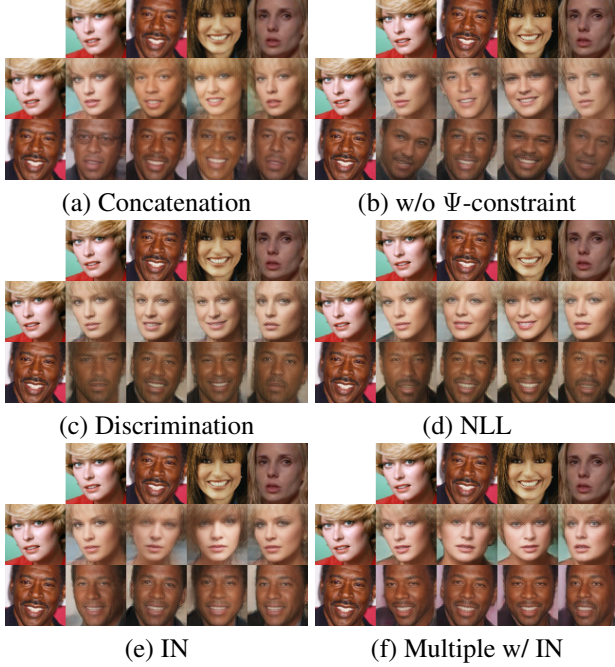


Figure 3. Ablation study of C-S DisMo. For each image, the content embedding is from the topmost image in the same column, and style embedding is from the leftmost image in the same row. A good disentanglement is that: horizontally, the style (identity) of the images is well maintained when the content (pose) varies, and vertically, the content of the images is well aligned when the style varies.

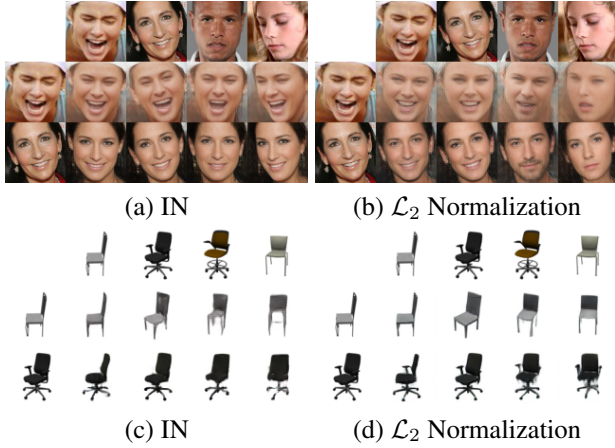


Figure 4. Comparison of the disentanglement with different normalizations. Instance Normalization (IN) achieves better results on CelebA, *e.g.*, the face identities are more alike. \mathcal{L}_2 normalization outperforms on Chairs, where the shapes of chairs are more consistent in each row.

neck capacity of content and style to help to avoid leakage. We introduce an information bottleneck given by

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

where C_s and C_c are the information capacity controlling the amount of information of the content and style, respectively.

During training, C_s and C_c increase linearly. The rate of increase is controlled by the increase steps and the maximum value. By controlling the increase rate, the content is forced to encode information first so that the learning process is more consistent with our assumptions.

Full objective. Therefore, our full objective is

$$w_{CS}\mathcal{L}_{CS} + w_{IB}\mathcal{L}_{IB} + w_{ID}\mathcal{L}_{ID}, \quad (7)$$

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

4. Experiments

In this section, we perform quantitative and qualitative experiments to evaluate our method. We test our method on several datasets: **Car3D** [28], **Chairs** [1] and **CelebA** [24]. For these three datasets, pose is the most dominant factor and encoded by content. For details of the datasets, please refer to the supplementary material.

Baselines. We choose several SOTA group-supervised C-S disentanglement benchmarks for comparisons: CycleVAE [12], DrNet [5] and Lord [6]. We select the only unsupervised C-S disentangled method Wu et al. [34]¹. We choose one VAE-based unsupervised disentanglement method: FactorVAE [14]. For FactorVAE, according to our definition of content and style, we manually traverse the latent space to select the factors related to pose as content and treat the other factors as style, for all these three datasets. More details for baselines are presented in the supplementary material.

4.1. Quantitative Experiments

We compare our method (Multiple C-S DisMo framework) with the baselines on Car3D, Chairs and CelebA.

Content Transfer Metric. To evaluate our method’s disentanglement ability, we follow the protocol of Gabbay and Hoshen [6] to measure the quality of content transfer by LPIPS [39]. Details are presented in Appendix A. The results are shown in Table 1. We achieve the best performance among the unsupervised methods, even though pseudo labels are provided for Wu et al. [34]. Our method significantly outperforms FactorVAE, which verifies the effectiveness of our formulation: simplifying the problem from disentangling factors to disentangling content and style. Furthermore, our method is comparable to or even better than the supervised ones.

Classification Metric. Classification accuracy is used to evaluate disentanglement in [5, 12, 6]. we follow the protocol of Jha et al. [12]. Low classification accuracy indicates

¹There is no open-sourced implementation for it. We modify <https://github.com/CompVis/vunet> and provide pseudo ground truth landmarks to the network. Thus it becomes semi-supervised actually.

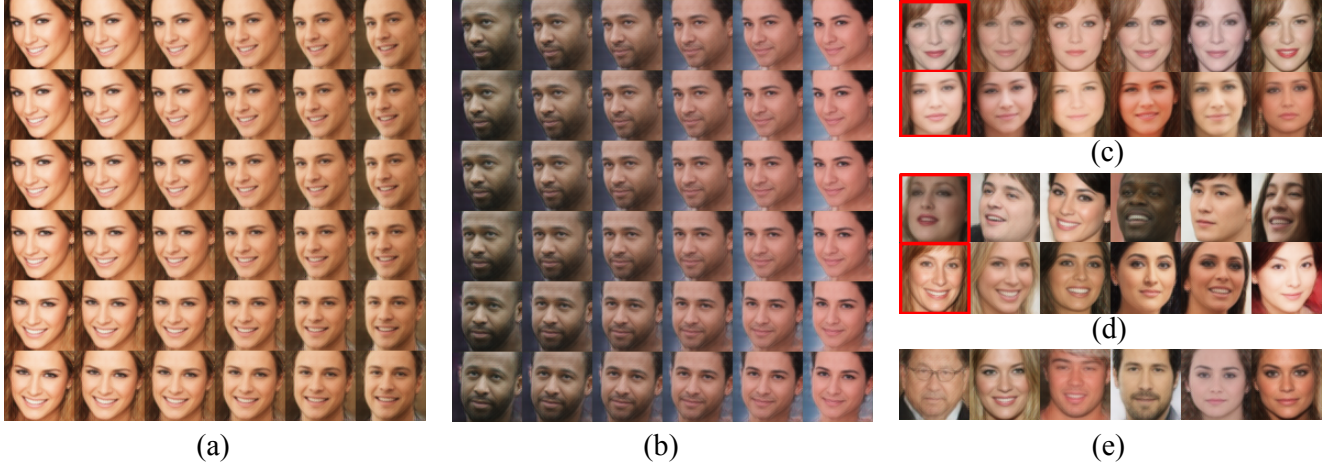


Figure 5. Demonstrations of the content and style space by interpolation (a & b) and retrieval (c-e).

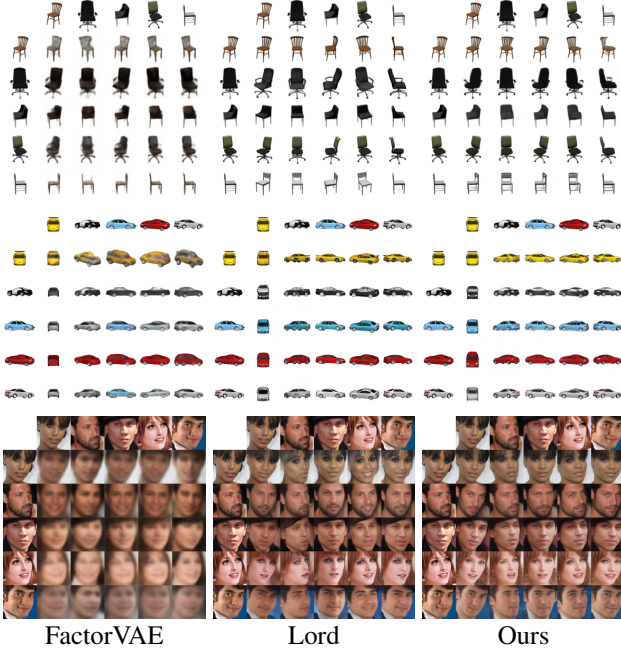


Figure 6. Comparison of visual analogy results on Chairs, Car3D and CelebA (from top to bottom). Zoom in for details.

small leakage between content and style. Without content annotations for CelebA, we regress the position of the facial landmarks from the style embeddings instead. The results are summarized in Table 2. Though without supervision, the performance of our method is still comparable to several other methods. We note that the classification metric may not be appropriate for disentanglement, which is also observed in Liu et al. [23]. The observation is that the classification metric is also influenced by information capacity and dimensions of embeddings. For FactorVAE, the poor reconstruction quality indicates that the content and style embeddings encode little information that can hardly be identified. The dimensions of the content and style embeddings of different methods vary from ten to hundreds, and a higher

Method	Supervision	Cars3D	Chairs	CelebA
DrNet [5]		0.146	0.294	0.221
Cycle-VAE [12]	✓	0.148	0.240	0.202
Lord [6]		0.089	0.121	0.163
FactorVAE [14]		0.190	0.287	0.369
Wu et al. [34] ¹	✗	—	—	0.185
Ours		0.082	0.190	0.161

Table 1. Performance comparison in content transfer metric (lower is better). For Wu et al. [34]¹, we provide pseudo facial landmarks, and there are no suitable landmarks for cars and chairs.

Method	Supervision	Cars3D		Chairs		CelebA	
		$s \rightarrow c$	$s \leftarrow c$	$s \rightarrow c$	$s \leftarrow c$	$R(s) \rightarrow c$	$s \leftarrow c$
DrNet [5]		0.27	0.03	0.06	0.01	4.99	0.00
Cycle-VAE [12]	✓	0.81	0.77	0.60	0.01	2.80	0.12
Lord [6]		0.03	0.09	0.02	0.01	4.42	0.01
FactorVAE [14]		0.07	0.01	0.14	0.01	5.34	0.00
Wu et al. [34] ¹	✗	—	—	—	—	5.42	0.11
Ours		0.33	0.24	0.66	0.05	4.15	0.05

Table 2. Classification accuracy of style labels from content codes ($s \leftarrow c$) and of content labels from style codes ($s \rightarrow c$) (lower is better). For Wu et al. [34], we provide pseudo ground truth landmarks. Note that the column ($R(s) \rightarrow c$) presents the error of face landmark regression from the style embeddings (higher is better).

dimension usually leads to easier classification.

4.2. Qualitative Experiments

Disentanglement & Alignment. In Figure 5 (a) and (b), we conduct linear interpolation to show the variation in the two embedding spaces. Horizontally, with the interpolated style embeddings, the identity (style) is changed smoothly while the pose (content) is well maintained. Vertically, the identity remains the same as the pose changes. We have the following observations: (i) The learned content and style spaces are continuous. (ii) Columns of the left and right figures share the same pose, suggesting that the learned content spaces are well aligned. (iii) Factors encoded by



Figure 7. Comparison of visual analogy results on Market-1501 dataset. Our method outperforms FactorVAE [14] significantly.

style is maintained when changing the content embeddings and vice versa, suggesting the good disentanglement.

We perform retrieval on the content and style latent spaces, respectively. As shown in Figure 5 (c) and (d), given a query image (labeled with a red box), its nearest neighbors in the content space share the same pose but have different identities, which reveals the content space is well aligned. To better identify the faces, we let the query’s nearest neighbors in the style space share the same pose, and the generated faces look very similar, revealing that the style is well maintained. As shown in Figure 5 (e), zero-valued content embedding result in a canonical view. As we assume that the pose distribution of faces is $\mathcal{N}(0, I)$, the canonical views are the most common pose in the dataset, and the zero-valued content embedding has the largest likelihood accordingly.

Visual Analogy & Comparison. Visual analogy [28] is to switch style and content embeddings for each pair. We show the visual analogy results of our method against FactorVAE (typical unsupervised baseline) and Lord (strongest supervised baseline) in Figure 6 on Chairs, Car3D, and CelebA. The results show that FactorVAE on all datasets is of poor generation quality and bad content transfer. On Cars3D, Lord’s results have artifacts (*e.g.*, third column), and its style embeddings could not encode the color information of the test images (*e.g.*, fourth row). On CelebA, the transfer result of Lord is not consistent, *e.g.*, the content embedding controls facial expression in the fifth column, while other content embeddings do not control expression. Our method achieves comparable pose transfer to Lord and maintains the identities of the images. Furthermore, we show our results on the Market-1501 [40] dataset in Figure 7, which demonstrates our method can disentangle the human pose and the appearance even though the skeletons have large variances. For more results (including on other datasets), please refer to the supplementary material.

Method	Content transfer metric ↓	Classification metric	
		$R(s) \rightarrow c \uparrow$	$s \leftarrow c \downarrow$
Single	0.204	3.03	0.06
Single w/ Disc	0.178	2.97	0.14
Single w/ NLL	0.171	2.98	0.09
Single w/ IN	0.166	3.46	0.04

Table 3. Ablation study for different solutions for Ψ -*constraint* on CelebA. Disc means discrimination-based solution.

4.3. Ablation Study

Choice of Ψ -constraint. Beside the qualitative experiment shown in Figure 4, we perform ablation study on CelebA to evaluate different solutions for Ψ -*constraint* introduced in Section 3.2. In this subsection, we do not use auxiliary loss functions. As shown in Table 3, all the solutions can achieve the SOTA performance in terms of content transfer metric, which means that the Ψ -*constraint* for content embeddings is crucial. This result further verifies that our definition is reasonable. For the classification metric, the results of discrimination-based and NLL-based solutions are relatively poor due to the reasons discussed in Section 4.1. The normalization-based solution achieves the best results on all the metrics. We believe that is because the normalization-based solution does not introduce an extra optimization term, which may hurt the optimization process and limit the expressive ability of embeddings.

Choice of size of embeddings. We also conduct experiments on the influence of the size of embeddings. We empirically set the size of style embedding d_s to 256 and the size of content embedding d_c to 128, which achieves good performance on all the datasets. Here, we demonstrate that we can also control the role of style by adjusting the size of the embeddings, as shown in Figure 9. For Figure 9 (a), the content embeddings contain the shape of the face, facial expression, and pose. For Figure 9 (b), the content embeddings contain the shape of the face and facial expression. For Figure 9 (c), which is the setting used in our paper, the content embeddings contain pose.

4.4. Unseen Images Inference

Our method can be generalized to the held-out data. A solution is to train two encoders to encode images to the content and style spaces respectively. We train a style encoder E_s and a content encoder E_c by minimizing

$$\mathcal{L}_E = \sum_{i=1}^N \|E_s(I_i) - s_i\|_1 + \|E_c(I_i) - c_i\|_1. \quad (8)$$

We apply our model trained on the CelebA dataset to faces collected by Wu et al. [33] including paintings and cartoon drawings. As shown in Figure 11, our method can be well generalized to unseen images from different domains.

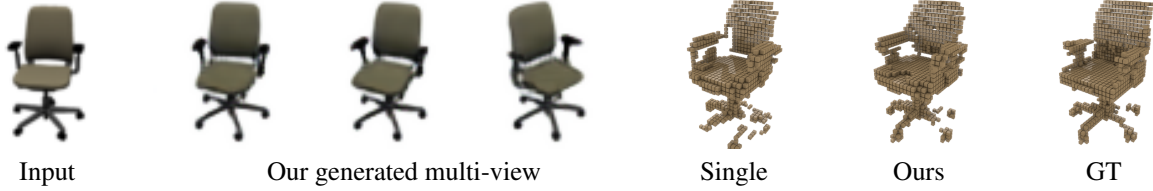


Figure 8. 3D reconstruction results on Chairs. Single: the object reconstructed by only Input. Ours: the object reconstructed from multi-view inputs generated by our method from Input. GT: the object reconstructed by the ground truth of multi-view inputs.

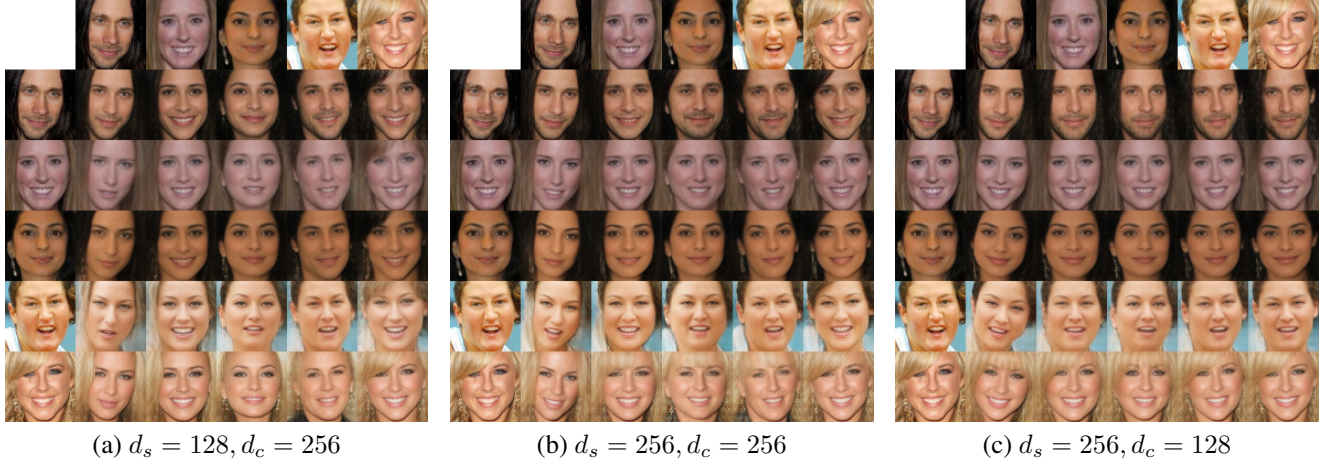


Figure 9. Study on the influence of size of embeddings. d_s is the size of style embedding and d_c is the size of content embedding.



Figure 10. Examples of translating shoes to edge (left column) and translating edges to shoes (right column). Triplet order (left to right) is: content, style, translation.

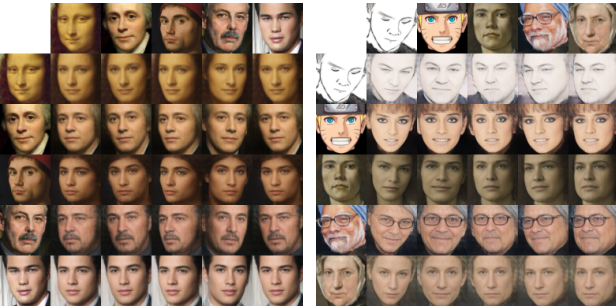


Figure 11. Inference for unseen images. Our method performs well on images from different domains: painting and cartoon.

4.5. Extension for Applications

In this work, we explore two applications of C-S disentanglement. For 3D reconstruction, single-view settings lack reliable 3D constraints [32]. Based on our disentangled representations, we can generate multi-view from a single view. On Chairs, we adopt Pix2Vox [37], a framework for

single-view, and multi-view 3D reconstruction to verify the effectiveness of our method. As shown in Figure 8, the 3D objects reconstructed from multi-view generated from our method are much better than those reconstructed from a single view, and even comparable to those reconstructed from ground-truth multi-view.

For domain translation, our method can work on the images merged from two domains without using any domain label. As shown in Figure 10, based on the disentangled content (edge structure) and style (texture), we can translate edge images into shoe images and vice versa.

5. Conclusion

We propose a definition for content and style and a problem formulation for unsupervised C-S disentanglement. Based on the formulation, C-S DisMo is proposed to assign different and independent roles to content and style when approximating the real data distributions. Our method outperforms other unsupervised approaches and achieves comparable or even better performance than the SOTA supervised methods. As for the limitation, we fail on datasets containing multiple categories with large appearance variation, *e.g.*, CIFAR-10 [18], which does not match our assumption. As for application, our method could be adopted to help downstream tasks, *e.g.*, domain translation, single-view 3D reconstruction, etc.

References

- [1] 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. 2, 4, 5
- [2] Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *AAAI*, 2018. 1, 2
- [3] 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*, 2018. 1, 2, 3, 4
- [4] Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *NeurIPS*, 2018. 1, 2
- [5] Emily L. Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. In *NeurIPS*, 2017. 1, 5, 6
- [6] Aviv Gabbay and Yedid Hoshen. Demystifying inter-class disentanglement. In *ICLR*, 2020. 1, 2, 5, 6
- [7] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016. 3
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4
- [9] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017. 1, 2
- [10] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 3
- [11] Xun Huang, Ming-Yu Liu, Serge J. Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 2
- [12] 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. 1, 2, 5, 6
- [13] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 3
- [14] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *ICML*, 2018. 1, 2, 5, 6, 7
- [15] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*, 2018. 4
- [16] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 2
- [17] Dmytro Kotovenko, Artsiom Sanakoyeu, Sabine Lang, and Bjorn Ommer. Content and style disentanglement for artistic style transfer. In *ICCV*, 2019. 1
- [18] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 8
- [19] Chuan Li and Michael Wand. Combining markov random fields and convolutional neural networks for image synthesis. In *CVPR*, 2016. 3
- [20] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. In *IJCAI*, 2017. 3
- [21] 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. 1
- [22] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *ICCV*, 2019. 2
- [23] Xiao Liu, Spyridon Thermos, Gabriele Valvano, Agisilaos Chartsias, Alison O’Neil, and Sotirios A. Tsaftaris. Metrics for exposing the biases of content-style disentanglement. *CoRR*, abs/2008.12378, 2020. 6
- [24] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 1, 5
- [25] 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. 1, 2
- [26] 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. 2
- [27] Mervin E Muller. A note on a method for generating points uniformly on n-dimensional spheres. *Communications of the ACM*, 2(4):19–20, 1959. 4
- [28] Scott E. Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. In *NeurIPS*, 2015. 5, 7
- [29] Xuanchi Ren, Haoran Li, Zijian Huang, and Qifeng Chen. Self-supervised dance video synthesis conditioned on music. In *ACM MM*, 2020. 3
- [30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 3
- [31] Attila Szabó, Qiyang Hu, Tiziano Portenier, Matthias Zwicker, and Paolo Favaro. Challenges in disentangling independent factors of variation. In *ICLRW*, 2018. 2
- [32] Fanzi Wu, Linchao Bao, Yajing Chen, Yonggen Ling, Yibing Song, Songnan Li, King Ng Ngan, and Wei Liu. Mvf-net: Multi-view 3d face morphable model regression. In *CVPR*, 2019. 8
- [33] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *CVPR*, 2020. 3, 7
- [34] Wayne Wu, Kaidi Cao, Cheng Li, Chen Qian, and Chen Change Loy. Disentangling content and style via unsupervised geometry distillation. In *ICLRW*, 2019. 1, 2, 3, 5, 6
- [35] Wayne Wu, Kaidi Cao, Cheng Li, Chen Qian, and Chen Change Loy. Transgaga: Geometry-aware unsupervised image-to-image translation. In *CVPR*, 2019. 1
- [36] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 4

- [37] 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. 8
- [38] 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. 1
- [39] 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. 5
- [40] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 7