DISENTANGLING THE SPATIAL STRUCTURE AND STYLE IN CONDITIONAL VAE

Ziye Zhang, Li Sun, Zhilin Zheng and Qingli Li

Shanghai Key Laboratory of Multidimensional Information Processing,
Key Laboratory of Advanced Theory and Application in Statistics and Data Science
East China Normal University, 200241 Shanghai, China

ABSTRACT
This paper aims to disentangle the latent space in cVAE into
the spatial structure and the style code, which are complement-
tary to each other, with one of them \( z_s \) being label relevant and
the other \( z_u \) irrelevant. The generator is built by a connected
encoder-decoder and a label condition mapping network. De-
pending on whether the label is related with the spatial struc-
ture, the output \( z_s \) from the condition mapping network is
used either as a style code or a spatial structure code. The
encoder provides the label irrelevant posterior from which \( z_u \)
is sampled. The decoder employs \( z_s \) and \( z_u \) in each layer by
adaptive normalization like SPADE or AdaIN. Extensive ex-
periments on two datasets with different types of labels show
the effectiveness of our method.

Index Terms— cVAE, GAN, disentanglement

1. INTRODUCTION

VAE [1] and GAN [2] are two powerful tools for image syn-
thesis. In GAN, the generator \( G(z) \) aims to mimic the data
distribution \( p_{\text{data}}(x) \) with an approximation \( p_G(z) \) by map-
ing the random noise \( z \) drawn from prior to the image-like
data. Meanwhile, GAN learns a discriminator \( D \) to distin-
guish the source of samples, either drawn from \( p_{\text{data}}(x) \) or
\( p_G(z) \). \( G \) and \( D \) are trained jointly in an adversarial man-
ner. VAE consists of a pair of connected encoder and de-
coder. The encoder \( q_\phi(z|x) \) maps the data \( x \) into a code \( z \),
and decoder \( p_\theta(x|z) \) transforms \( z \) into image domain and tries
to reconstruct \( x \). \( q_\phi(z|x) \) is to be simple, e.g., close to the standard
Gaussian prior \( N(0, I) \) based on the KL divergence metric.

Compared to GAN, VAE tends to generate blurry im-
ages, since \( q_\phi(z|x) \) is too simple to capture the true posterior,
known as ”posterior collapse”. But it is easier to train. While
GAN’s optimization is unstable, hence many works try to
stabilize its training [3][5]. Moreover, VAE explicitly mod-
els each dimension of \( z \) as independent Gaussian, so it can
disentangle the factors in unsupervised way [6][7]. To fully
exploit the advantage from both of them, VAE and GAN can
be combined into VAE-GAN [8], in which the encoder and
decoder in VAE forms the generator, and it employs a dis-

...
spatial related or not, and posteriors in both of them are constrained in the same way, close to the prior \( N(0, I) \), thus the code sampling from it is label irrelevant. Particularly, the spatial structure code, either sampling from the posterior or mapping from the label, is applied into the encoder and decoder through spatially-adaptive normalization (SPADE) [15]. On the other hand, the style code is incorporated into VAE by adaptive instance normalization (AdaIN) [16][17]. To improve the quality of the synthesis, we add a discriminator like cVAE-GAN [12], and extend the fake data.

Our contribution lies in following aspect. First, we propose a simple, flexible method to disentangle the spatial structure and style code for image synthesis. We only require one of them is label dependent, and it is given to the generator as the condition. The other is fully unsupervised and becomes label irrelevant after training. Second, by applying the adaptive normalization based on both the style and the spatial structure code, our model improves the disentangling performance. We carry out experiments on two types of datasets to prove the effectiveness of the method.

2. PROPOSED METHOD

We now give the details about the proposed method. Fig. 1 shows the overall architecture. The generator \( G \) consists of \( Enc, Dec \) and a small condition mapping network \( f \). \( Enc \) specifies the spatial structure preserving posterior \( N(\mu, \sigma) \) which is assumed to be label irrelevant, and constrained with prior \( N(0, I) \) by KL divergence. \( Dec \) exploits the spatial structure code \( z_u \sim N(\mu, \sigma) \) with SPADE, and the style code \( w \) given by \( f \) with AdaIN. The connected \( Enc \) and \( Dec \) are trained by the reconstruction loss \( L_{rec} \). The discriminator \( D \) helps \( G \) synthesizing the high quality image by providing the adversarial GAN loss \( L_{adv}^{D/G} \).

![Fig. 1. The proposed architecture on FaceScrub, in which ID labels have few spatial cues, and they are mapped into style codes. \( Dec \) exploits the spatial structure code \( z_u \sim N(\mu, \sigma) \) with SPADE, and the style code \( w \) given by \( f \) with AdaIN. Note that the best architecture is different on 3D chair.](image)

2.1. Problem formulation

Supposing the image is \( x \) and its label is \( c \), the goal of cVAE is to maximize the ELBO defined in [1], so that the data log-likelihood \( \log p(x) \) can be maximized. As is described in [14], the key idea is to split the latent code \( z \) into separate codes, the label relevant \( z_u \) and irrelevant \( z_s \). Here \( \phi, \psi \) and \( \theta \) correspond to the model parameters in \( Enc, Dec \) and \( f \), respectively. \( D_{KL} \) indicates the KL divergence between two distributions. Note that there are three terms in [1]. The first one is the negative reconstruction error. The second and third terms are the regularization which pushes the \( q_\phi(z_u|x,c) \) and \( q_\psi(z_s|c) \) to their priors \( p(z_u) \) and \( p(z_s) \), respectively. In practice, we assume that \( z_s \) is deterministic, which means \( p(z_s) \) and \( q_\psi(z_s|c) \) are both dirac \( \delta \) function. Hence the third term is strictly required to be 0, thus can be ignored.

\[
\log p(x) \geq E_{q_\phi(z_u|x,c),q_\psi(z_s|x)}[\log p_\theta(x|z_u, z_s)]
- D_{KL}(q_\phi(z_u|x,c)||p(z_u))
- D_{KL}(q_\psi(z_s|c)||p(z_s))
\]

2.2. Details about the network

2.2.1. Conditional label mapping network \( f \)

As is illustrated in Fig. 1, the input of \( f \) is a conditional label \( c \), which indicates the category of \( x \), usually expressed in one-hot format. Like [17], we first use several fully-connected layers to map \( c \) into an embedding code \( w \), which is later used by both the \( Enc \) and \( Dec \) based on the adaptive normalization module. Here the output \( w \) is regarded as the label relevant code \( z_u \) which is \( w = z_u \), and it is treated either as a spatial structure preserving feature map with its dimension \( H \times W \times 1 \), or a style feature with its size \( 1 \times 1 \times C \). We make the choice based on whether \( c \) is directly related with the spatial structure. Actually, we try both cases on two different datasets, 3D chair [18] and FaceScrub [19]. The details are given in the experiments section.

2.2.2. Encoder and Decoder

cVAE has a pair of connected \( Enc \) and \( Dec \). The \( Enc \) takes the image and label pair \( x, c \) and maps it into a posterior probability, which is assumed to have the Gaussian form \( p(z_u|x,c) = N(z_u|x,c; \mu, \sigma) \). Here \( \mu \) and \( \sigma \) are the mean and standard deviation, which are two outputs from \( Enc \) depending on \( x \). A code \( z_u \sim p(z_u|x,c) \) is sampled from it to synthesize the reconstruction of \( x \). Note that in VAE, there is a prior assumption on \( z_u \), which is \( p(z_u) = N(z_u; 0, I) \). During the optimization, \( D_{KL}(p(z_u|x,c)||p(z_u)) \) is considered. In other words, the \( Enc \) tries to map \( x \) from various classes into the same prior. Therefore \( z_u \) tends to be label irrelevant in cVAE.

In our scheme, we also have two choices on the output of the \( Enc \), which are closely related to output \( z_s \) (or \( w \)) from the condition mapping network \( f \). Particularly, \( z_u \) can be assumed as either a spatial structure code with its size \( H \times W \times 1 \) or a style code with the dimension \( 1 \times 1 \times C \). In the first case, the spatial structure is kept from \( x \) so there should be no fully-connected layers in the \( Enc \). In the second case, \( z_u \) is
formulated to capture the feature channel style based on the global average pooling. To make \( z_s \) complementary to \( z_u \), we intentionally design it so that one of \( z_u \) or \( z_s \) is the style code and the other is the spatial structure code. In the experiments, we demonstrate that different dimension settings for \( z_s \) and \( z_u \) improve disentangling performances.

The \( \text{Dec} \) takes the code \( z_s \) and \( z_u \) and tries to reconstruct \( x \). Traditionally, all inputs are directly concatenated in the channel dimension, then fed into the \( \text{Dec} \) as its input. However, this simple strategy does not emphasize the difference between \( z_s \) and \( z_u \), which definitely degrades the disentangling quality. Inspired by two adaptive normalization structure, AdaIN [17] and SPAD [15], we find they are suitable for our task. As is shown in (2), \( h^{(l)} \) is processed by SPADE and AdaIN and disentangling quality. A model makes a simple change on cVAE-GAN, incorporates label condition into \( \text{Enc} \) and \( \text{Dec} \) by exploiting label relevant code \( z_u \) with AdaIN, and the spatial code \( z_s \) with SPAD.

### 3. EXPERIMENTS

#### 3.1. Experimental Setup

**Datasets.** We conduct experiments on two datasets, including the 3D chair [18] and the FaceScrub [19]. The 3D chair depicts a wide variety of chairs in 62 different chosen azimuth angles. Images are resized to the fixed size 64 \( \times \) 64. The FaceScrub contains 107k facial images from 530 different IDs. These faces are cropped by the detectors [20], and they are aligned based on the facial landmarks [21]. The detected cropped images are in 128 \( \times \) 128.

**Evaluation metrics.** We adopt three metrics for quantitative analysis. (1) Classification Accuracy (Acc) reflects the condition conformity of the generated images. We use models of ResNet-50 [22] trained on these two datasets for evaluating. (2) Fréchet Inception Distance (FID) [23] measures the distance between distributions of the synthesized and the real images, thus the lower, the better. (3) Mutual Information (MI) between label irrelevant code \( z_u \) and original label \( c \). If label relevant and irrelevant variables are disentangled well, the mutual information value \( I(z_u; c) \) should be small:

\[
I(z_u; c) = \mathbb{E}_{q(z_u|c)p(c)} \log \frac{q(z_u|c)}{q(z_u)} = \frac{1}{C} \sum_c \mathbb{E}_{q(z_u|c)} \log \frac{q(z_u|c)}{q(z_u)}
\]

\( q(z_u|c) \) and \( q(z_u) \) are hard to be computed directly, but can be approximated with Monte Carlo simulation.

#### 3.2. Results

**3D chair.** The label of 3D chair is azimuth angle. The model can generate chair with the specified azimuth, and the output image should preserve original chair style. Here, the spatial structure code is label relevant and the style code irrelevant. Different from Fig.1, \( \text{Dec} \) adopts the style code \( z_u \sim N(\mu, \sigma) \) with AdaIN, and the spatial code \( z_s \) with SPAD.

We compare our method with cVAE-GAN, and other four models for ablation study. The first model S1 makes a simple change on cVAE-GAN, incorporates label condition into \( \text{Enc} \) and \( \text{Dec} \) by exploiting label relevant code \( z_u \) with AdaIN. S2 uses SPAD to process \( z_u \), which is reshaped to \( H \times W \times \times 1 \). S3 is similar with our proposed structure except the embedding code \( z_s \) is applied with AdaIN. S4 manipulates spatial structure code \( z_s \) with AdaIN, and style code \( z_u \) with SPAD.

**Facades.** FaceScrub provides 1D labels. The model is supposed to output facial image with specified ID but preserve label irrelevant information, like pose, expression, from the input image. It is obviously that the label relevant code controls the style and the label irrelevant code specifies spatial related information. Fig.1 shows our proposed method. Also, we compare our method with cVAE-GAN, and choose S1 and S3 described previously for ablation study.
Table 1. The comparison network structures on two datasets. For \( z_s \), two choices are AdaIN and SPADE. For \( z_u \), it can also be given to \( Dec \) directly. Note that we do not list and compare all possible ways. But these are typical ones.

|          | 3D chair | Facescrub |
|----------|----------|-----------|
| \( z_{\text{S}} \) | cVAE-GAN | S1 | S2 | S3 | S4 | Proposed |
| concat   | AdaIN    | SPADE | AdaIN | AdaIN | SPADE | AdaIN |
| \( z_{\text{U}} \) | Dec input | Dec input | Dec input | AdaIN | SPADE | Dec input |

Table 1 lists the comparison structures in our experiments on two datasets. “Dec inputs” indicates that the code is taken as \( Dec \)’s input. “AdaIN” or “SPADE” means the corresponding code is incorporated into \( Dec \) by this operation.

Fig. 2. Exchanged images on 3D chair are shown here.

Fig. 3. Reconstructed and exchanged images on FaceScrub.

3.3. Analysis

We choose one specific label to generate images. All the images we analyze here are exchanged images. For 3D chair, the input label is “24” and for FaceScrub the label is “Anne Hathaway”. The generated images from different models on 3D chair and FaceScrub are presented in Fig.2 and Fig.3, respectively. In both Fig.2 and Fig.3, images generated from our proposed method achieves the best performance.

Table 2. MI and Acc on 3D chair from different models.

|         | MI   | Acc  |
|---------|------|------|
| cVAE-GAN | 3.777 | 0.124 |
| S1      | 3.733 | 0.573 |
| S2      | 3.765 | 0.608 |
| S3      | 3.775 | 0.511 |
| S4      | 3.774 | 0.404 |
| Proposed method | **3.750** | **0.623** |

Table 3. Acc and FID on FaceScrub from different models.

|         | Acc  | FID  |
|---------|------|------|
| cVAE-GAN | 0.072 | 83.05 |
| S1      | 0.444 | 80.37 |
| S3      | 0.498 | 52.46 |
| Proposed method | **0.632** | **50.14** |

4. CONCLUSION

We propose a latent space disentangling algorithm for conditional image synthesis in cVAE. Our method divides the latent code into label relevant and irrelevant parts. One of them preserves the spatial structure, and the other is the style code. These two types of codes are applied into cVAE by different adaptive normalization schemes. Together with a discriminator in the pixel domain, our model can generate high quality images, and achieve the disentangling performance.
5. REFERENCES

[1] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra, “Stochastic backpropagation and approximate inference in deep generative models,” arXiv preprint arXiv:1401.4082, 2014.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, “Generative adversarial nets,” in Advances in neural information processing systems, 2014, pp. 2672–2680.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou, “Wasserstein generative adversarial networks,” in International Conference on Machine Learning, 2017, pp. 214–223.

[4] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville, “Improved training of wasserstein gans,” in Advances in Neural Information Processing Systems, 2017, pp. 5767–5777.

[5] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida, “Spectral normalization for generative adversarial networks,” arXiv preprint arXiv:1802.05957, 2018.

[6] 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,” ICLR, vol. 2, no. 5, pp. 6, 2017.

[7] Tian Qi Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud, “Isolating sources of disentanglement in variational autoencoders,” in Advances in Neural Information Processing Systems, 2018, pp. 2610–2620.

[8] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther, “Autoencoding beyond pixels using a learned similarity metric,” arXiv preprint arXiv:1512.09300, 2015.

[9] Mehdi Mirza and Simon Osindero, “Conditional generative adversarial nets,” arXiv preprint arXiv:1411.1784, 2014.

[10] Takeru Miyato and Masanori Koyama, “cgans with projection discriminator,” arXiv preprint arXiv:1802.05637, 2018.

[11] Kihyuk Sohn, Honglak Lee, and Xinchen Yan, “Learning structured output representation using deep conditional generative models,” in Advances in neural information processing systems, 2015, pp. 3483–3491.

[12] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua, “Cvae-gan: fine-grained image generation through asymmetric training,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2745–2754.

[13] Jack Klys, Jake Snell, and Richard Zemel, “Learning latent subspaces in variational autoencoders,” in Advances in Neural Information Processing Systems, 2018, pp. 6442–6454.

[14] Zhilin Zheng and Li Sun, “Disentangling latent space for vae by label relevant/irrelevant dimensions,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 12192–12201.

[15] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu, “Semantic image synthesis with spatially-adaptive normalization,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2337–2346.

[16] Xun Huang and Serge Belongie, “Arbitrary style transfer in real-time with adaptive instance normalization,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1501–1510.

[17] Tero Karras, Samuli Laine, and Timo Aila, “A style-based generator architecture for generative adversarial networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410.

[18] 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 Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 3762–3769.

[19] Hong-Wei Ng and Stefan Winkler, “A data-driven approach to cleaning large face datasets,” in 2014 IEEE International Conference on Image Processing (ICIP). IEEE, 2014, pp. 343–347.

[20] Dong Chen, Shaoqing Ren, Yichen Wei, Xudong Cao, and Jian Sun, “Joint cascade face detection and alignment,” in European Conference on Computer Vision. Springer, 2014, pp. 109–122.

[21] Xuehan Xiong and Fernando De la Torre, “Supervised descent method and its applications to face alignment,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 532–539.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” in Advances in Neural Information Processing Systems, 2017, pp. 6626–6637.