Generative Adversarial Networks for Diverse and Explainable Text-to-Image Generation

Zhang, Zhenxing

DOI:
10.33612/diss.507581028

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2023

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
It has been observed that the latent space of a GAN contributing to diversity contains a wide range of semantics, playing a significant role in the image-generation process. Although many approaches have been proposed to interpret the latent space of a GAN, there still is a lack of understanding about how a cGAN transforms the latent space to the distribution of generated images. In this chapter, we concentrate on studying the following issue: How to exploit the latent space of a conditional text-to-image GAN model to edit a synthesized image? More specifically, we present a novel algorithm which identifies semantically-understandable directions in the latent space of a conditional text-to-image GAN architecture by performing the independent component analysis algorithm under an additional orthogonality constraint on the pretrained weight values of the generator. The captured directions are not only independent but also orthogonal. Furthermore, we develop a background-flattening loss (BFL) to improve the background appearance in an edited image. Also, we mathematically analyze the relationships between Semantic Factorization (SeFa), GANSpace and regular PCA and show that they typically achieve almost the identical results when sampling enough data for GANSpace. Experimental results on the recent single-stage text-to-image GAN models pre-trained on three benchmark data sets demonstrate that our proposed approach is able to derive various interpretable semantic properties and provide a more precise control over the latent space than PCA, validating the effectiveness of our presented techniques.

This chapter is based on
5.1 INTRODUCTION

Existing text-to-image approaches mostly concentrate on improving both image quality and semantic relevance but ignore the explainability of a conditional text-to-image GAN model that plays an import role in facilitating its applicability in practice. This chapter intends to explain a conditional text-to-image GAN model from the latent-space point of view, since latent space contributes to diversity.

Recent works [34], [94], [95] reveal that there exists a wide range of meaningful semantic factors in the latent space of a GAN, such as facial attributes and head poses for face synthesis [94] and layout for scene generation [120]. These semantically-understandable control directions can be utilized for disentangled image editing, like semantic face editing [94] and scene manipulation [120]. By moving the latent code of a synthetic sample towards and backwards the direction, we are able to vary the desired attribute while keeping other image contents unchanged. That is to say, given a successful latent code, we can derive a wealth of similar but semantically-diverse pleasing images via latent-space navigation. To better facilitate the application of text-to-image synthesis, we need to address the question: How to identify useful control directions in the latent space of a conditional text-to-image GAN model? While current approaches mainly focus on studying the latent space of a GAN, there still is a lack of understanding of the relationship between the latent space of a cGAN and the explainable semantic space in which a synthetic sample is embedded.

In this chapter, we present a novel algorithm to capture the interpretable latent-space semantic properties of a conditional text-to-image GAN model. Considering the fact that identified directions denote different semantic factors of the edited object, e.g., pose and smile for the face model, we argue that these vectors should be fully independent rather than just uncorrelated. Based on recent studies [34], [95], we assume that the pretrained weights of a conditional text-to-image GAN architecture contain a set of useful directions. In fact, the initial linear layer projects the latent vector to the visual feature map, where a latent space is transformed into another space and ultimately into an output image. To acquire both independent and orthogonal components, we introduce the independent component analysis (ICA) algorithm under an additional orthogonality constraint to investigate the pretrained weight matrix of the first dense layer. In addition, we mathematically show that Semantic Factorization (SeFa) [95], GANSpace [34] and regular PCA [115] typically achieve almost the identical results when sampling enough data for GANSpace. Furthermore, we develop a Background-Flattening Loss (BFL) to improve the background appearance in an edited sample. Multiple interesting latent-space directions found by our presented algorithm are visualized in Fig. 5.1. We expect that our proposed semantic-discovery method can provide valuable insight into the relationship between latent vectors and
This bird is brown and white in color, with a brown beak.

The person has wavy hair, and high cheekbones. She is young and wears necklace, heavy makeup, and lipstick.

Figure 5.1: Interpretable latent-space directions identified in DiverGAN that was pre-trained on the CUB bird [108] (left side) and Multi-Modal CelebA-HQ [116] (right side) data sets. For each set of pictures, the middle column is the original image based on a Good latent code, while the samples on the left and right of it are the output by freezing the textual description and moving the latent vector backward and forward from the center, over the axis discovered by our proposed algorithm.

obtained image variations. This will improve the explainability of the latent space of a conditional text-to-image GAN model.

We evaluate our proposed algorithm on the recent single-stage text-to-image GAN models, i.e., DF-GAN [105], DTGAN and DiverGAN. We carry out a serious of experiments on three popular text-to-image data sets, i.e., the CUB bird [108], MSCOCO [63] and Multi-Modal CelebA-HQ [116] data sets. Experimental results show that our presented semantic-discovery approach can lead to a more precise control over the latent space of a conditional text-to-image GAN model than PCA, which validates the effectiveness of our method. The contributions of this chapter can be summarized as follows:

- We mathematically analyze the relationships between SeFa [95], GANSpace [34] and regular PCA.
- We introduce the ICA algorithm under an additional orthogonality constraint to identify meaningful attributes in the latent space of a conditional text-to-image GAN model.
- We develop a Background-Flattening Loss (BFL) to improve the background appearance in an edited sample.
5.2 RELATED WORK

5.2.1 Study on the latent space of a GAN

Recent studies [34], [94], [95] on a GAN reveal that a latent space possess a variety of semantically-understandable information, e.g., pose and smile for the face data set, which plays a vital role in the disentangled sample manipulation. We are able to realistically edit a generated image by moving its latent vector towards the direction corresponding to the desired attribute. Several methods have been proposed to capture interpretable semantic factors and mainly fall into two types: (1) unsupervised models and (2) supervised approaches.

Supervised latent-space manipulation. Shen et al. [94] developed a framework termed as InterfaceGAN where labeled samples (e.g., gender and age) are utilized to train a linear SVM and the acquired SVM boundaries lead to the meaningful manipulation of the facial attributes. Goetschalckx et al. [29] proposed GANalyze applying an accessor module to optimize the training process while learning the latent-space directions as the desired cognitive semantics.

Unsupervised latent-space manipulation. Voynov et al. [107] introduced a matrix and a classifier to identify interpretable latent-space directions in an unsupervised fashion. Jahanian et al. [44] studied the attributes concerning color transformations and camera movements by operating source pictures. Häkkinen et al. [34] designed a novel pipeline named GANSpace, which performed PCA [115] on a series of collected latent vectors and employed obtained principal components as the meaningful directions in the latent space. Peebles et al. [109] presented the Hessian Penalty, a regularization term for the unsupervised discovery of useful semantic factors. Wang et al. [109] developed Hijack-GAN introducing an iterative scheme to control the image-generation process. Shen et al. [95] proposed Semantic Factorization (SeFa) which directly decomposed the weight matrix of a well-trained GAN model for semantic image editing. This chapter aims to identify controllable directions in the latent space of a conditional text-to-image GAN model.

5.3 PROPOSED METHODOLOGY

In this section, we mathematically show that Semantic Factorization (SeFa) [95] approximately identifies the principal components, as PCA does. Furthermore, we propose a simple but effective technique to capture semantically-interpretable latent-space directions for a conditional text-to-image GAN model. To optimize the edited sample, the background-flattening trick is presented to fine-tune the background appearance.
5.3 Proposed methodology

5.3.1 Problem statement

It has been widely observed [34], [94], [95] that the latent space of a GAN incorporates certain semantic information, like pose and size for the CUB bird data set. Suppose we have a Good latent code $z_g$ that contributes to a successful generated sample and a well-trained generator $G(z, (w, s))$ that can yield dissimilar and semantically consistent pictures according to a single textual descriptions and different injected noise, we target to manipulate the semantic factor of a successful synthesized sample via latent-space navigation. To this end, we need to identify a series of semantically-understandable latent-space directions $N = (n_1, n_2, \ldots, n_k)$, where $n_i \in \mathbb{R}^l$ for all $i \in 1, 2, \ldots, k$. After that, the attribute of the high-quality sample generated by $z_g$ can be varied by editing $z_g$ with $z_{ge} = z_g + \alpha n$, where $\alpha$ denotes the manipulation intensity and $n \in \mathbb{R}^l$ is the direction corresponding to the desired property.

5.3.2 Analyzing the correspondences between SeFa, GANSpace and PCA

We attempt to discuss the relationships between SeFa [95] and GANSpace [34], since they both introduce an algorithmically simple but surprisingly effective technique to derive semantically-understandable directions. Specifically, GANSpace collects a set of latent codes and conducts PCA on them to identify the significant latent-space directions. SeFa proposes to directly decompose the pretrained weights for semantic image editing. Mathematically, SeFa is formulated as:

$$A^T A n_i - \lambda_i n_i = 0 \quad (5.1)$$

where $A \in \mathbb{R}^{d \times l}$ is the weight matrix of the first transformation step in the generator and $\{n_i\}_{i=1}^{k}$ indicate $k$ most meaningful directions. The solutions to Equation 5.1 correspond to the eigenvectors of $A^T A$ with respect to the $k$ largest eigenvalues. $A$ is usually normalized by L2 norm when implementing SeFa. The formulation of SeFa can almost be perceived as PCA on $A$, since the results of PCA are the eigen vectors of the covariance matrix $C_A$ associated with $A$ and $C_A$ is similar to $A^T A$. Specifically, $C_A$ is denoted as:

$$C_A = \frac{1}{d-1} (A - <A>)^T (A - <A>) \quad (5.2)$$

where $<A>$ represents the mean from each column of $A$ and $C_A$ is the covariance matrix of $A$. The difference between regular PCA and SeFa is located in the normalization of $A$. We therefore argue that SeFa is approximately equivalent to regular PCA on the pretrained weights. That is to say, GANSpace and SeFa perform PCA on the latent vectors and the pretrained weights, respectively.
5.3.3 Independent component analysis for semantic discovery in the latent space

It has been observed that the pretrained weights of a standard GAN contain semantically-useful information. We can capture the meaningful latent-space directions in an unsupervised manner by exploiting the well-trained weights of a generator. A conditional text-to-image GAN generator typically leverages a dense layer to transform a latent code into a visual feature map, where a latent space is projected to another space and ultimately into an output image. We therefore make the assumption that there exists a wealth of semantics in the initial fully-connected weight matrix of a conditional text-to-image GAN model, due to the linguistic content of the text. We aim at presenting a simple but effective algorithm extracting the main patterns of the pretrained weights as the interpretable latent-space directions. More specifically, we hypothesize that when given the pretrained weight matrix $A$ of the first linear layer of $G(z, (w, s))$, we can obtain a suite of $k$ meaningful semantic factors $N = (n_1, n_2, \cdots, n_k)$ by processing the weight matrix $A$. Mathematically,

$$N = f(A)$$

where $f(\cdot)$ is the function for latent semantic discovery. These acquired semantics should denote different attributes of the generated image. For example, $n_1$ represents pose, $n_2$ represents smile and $n_3$ represents gender for the face data set. To better manipulate the image-generation process, we argue that these components should be fully independent rather than just uncorrelated (orthogonal). However, when employing PCA as $f(\cdot)$ to discover the controllable latent-space directions, the obtained principal components are only uncorrelated, but not independent. Meanwhile, PCA is optimal for Gaussian data only \cite{1}, while the pretrained weight matrix $A$ is not guaranteed to be Gaussian. Here, we propose to utilize independent component analysis (ICA) to identify useful latent-space semantics for a conditional text-to-image GAN model.

The goal of ICA is to describe a $M \times L$ data matrix $X$ in terms of independent components. It is denoted as:

$$X = BS$$

where $B$ is a $M \times T$ mixing matrix and $S$ is a $T \times L$ source matrix consisting of $T$ independent components.

ICA is commonly viewed as a more powerful tool than PCA \cite{17}, since it is able to make use of higher-order statistical information incorporating a variety of significant features. Furthermore, ICA is adequate for analyzing non-Gaussian data. To maximize both the independence and the orthogonality between the directions, i.e., $n_1, n_2, \cdots, n_k$, we apply a fast ICA under an additional orthogonality constraint \cite{1} to directly decompose the pretrained weight matrix to derive
the meaningful directions in the latent space. The obtained vectors are therefore not only independent but also orthogonal. We expect that the components can lead to a precise control over the latent space of DiverGAN.

5.3.4 Background flattening

A movement along an effective direction in the latent space should not only accurately change the desired attribute, but also maintain other image content, e.g., the background. However, when applying existing semantic-discovery methods even our introduced algorithm on a conditional text-to-image GAN model, we find that the background appearance in an edited sample usually varies along with the target attribute. To overcome this issue, we develop a Background-Flattening Loss (BFL) to fine-tune the acquired directions to improve the background appearance. This loss is defined by using both low-level pixels and high-level features, ensuring that the background is optimized and other image contents are preserved. Specifically, it is denoted as:

$$L_{flatten}(x_1, x_2) = ||x_1 - x_2||_1 + L_{LPIPS}(x_1, x_2)$$  \hspace{1cm} (5.5)

where $x_1, x_2$ refer to a source sample and an edited sample, respectively. We leverage the Adam algorithm [49] to optimize the acquired semantically-understandable latent-space directions.

We empirically discover that we are able to employ our proposed BFL to remove the patterns representing the background. To be specific, we can obtain a sample with a white background by increasing the distance (i.e., the BFL) between samples generated by different directions, since the white background and the black background will lead to the maximum loss values. After that, to remove the background, we take the white-background sample as the source image while reducing the distance between the source sample and the edited samples.

5.4 EXPERIMENTS

We evaluate our proposed approaches on a wide range of conditional text-to-image GAN models to identify meaningful latent-space directions. We perform a set of experiments on three benchmark data sets, i.e., the CUB bird [108], MSCOCO [63] and Multi-Modal CelebA-HQ [116] data sets. We show the principal components derived by SeFa [95], GANSpace [34] and regular PCA. After that, we visualize the interpretable semantic properties captured by our introduced algorithm in DF-GAN [105] and DTGAN. Afterwards, we qualitatively compare our method with existing unsupervised alternatives while performing a human test in order to verify its effectiveness. Subsequently, to evaluate the presented
background-flattening loss (BFL), we use the BFL to fine-tune the latent-space directions obtained by our approach on the CUB bird data set.

5.4.1 Experimental settings

Implementation details. We apply our presented semantic-discovery algorithm on the recent single-stage text-to-image GAN models, i.e., DF-GAN [105], DTGAN and DiverGAN. We set the manipulation intensity $\alpha$ to 3 for SeFa [95] and our proposed algorithm. The scalar parameter for GANSpace [34] is set to 20 on the CUB bird data set and 9 on the MSCOCO data set, respectively. We employ the Adam optimizer with $\beta = (0.0, 0.9)$ to fine-tune the identified latent-space directions. We set the learning rate to 0.0001. Our methods are implemented by PyTorch [81]. We conduct all the experiments on a single NVIDIA Tesla V100 GPU (32 GB memory).

5.4.2 Comparison between SeFa, GANSpace and PCA

Fig. 5.2 plots the latent-code manipulation results of SeFa [95], GANSpace [34] and regular PCA in DiverGAN on the CUB bird and MSCOCO data sets. We discover that these three approaches derive almost the identical directions although for some components (e.g., 4th principal component) the negative and the positive side is reversed, supporting our claim in Section 5.3.2. Note that GANSpace is implemented by leveraging the first dense layer of DiverGAN to collect 10,000 sets of feature maps while performing PCA on them to obtain the principal components as useful attributes. Additionally, we adjust the max manipulation intensity (i.e., $\alpha$ in Section 5.3.1) to 20 on the CUB bird data set and 9 on the MSCOCO data set, respectively. The above analysis suggests that when enough data is sampled, SeFa is similar to GANSpace for DiverGAN.

5.4.3 Results on DF-GAN and DTGAN

Results on DF-GAN. As described in Section 5.3.3, our proposed algorithm is able to discover a series of meaningful directions in the latent space of a conditional text-to-image GAN model. We evaluate our approach on the generator of DF-GAN pre-trained on the CUB bird and Multi-Modal CelebA-HQ data sets. Fig. 5.3 shows the semantically-understandable latent-space directions captured in DF-GAN. We can see that our presented method can effectively control multiple attributes, i.e., L/R orientation, size, position and branch for the bird data set and pose and hair for the face data set. It can also be observed that when altering one semantic, other image contents usually remain unchanged, which confirms the effectiveness of our approach. However, we find that our algorithm does not
Figure 5.2: Visualization of individual components within the latent codes discovered in DiverGAN, for (1) SeFa [95], (2) GANSpace [34] and (3) regular PCA. The original source image is in the left column (2 examples, a and b). For each principal component (pc1-pc4), example images from the negative and the positive side of its axis are shown.
Figure 5.3: Diverse interpretable directions found in DF-GAN [105], which was trained on (1) the CUB bird (four top rows) and (2) Multi-Modal CelebA-HQ data sets (four bottom rows). These semantics can be used to edit a synthetic sample.
5.4 experiments

identify some useful properties, e.g., background for the bird data set and smile for the face data set. This may be due to model or data-related limitations.

**Results on DTGAN.** We also utilize our proposed algorithm to interpret the latent space of DTGAN that was pre-trained on the CUB bird and Multi-Modal CelebA-HQ data sets. The obtained meaningful latent-space directions in DTGAN are depicted in Fig. 5.4. We can see that our method finds multiple interesting semantics, e.g., L/R orientation, size, pose and hair, similar to the useful latent-space directions captured in DF-GAN. The directions identified by our method can be employed to accurately edit a generated image. It qualitatively demonstrates that our approach provide a precise control over the latent space of a conditional text-to-image GAN model.

5.4.4 Comparison with unsupervised methods

**Qualitative results.** For qualitative comparison, we visualize the meaningful directions identified by our proposed algorithm and SeFa on the CUB bird and Multi-Modal CelebA-HQ data sets in Fig. 5.5. We can tell that our method is able to derive several fine-grained semantics corresponding to rotation, background and size for the bird model and pose, hair and smile for the face model, validating its effectiveness. Meanwhile, our approach leads to a more powerful control over the latent codes than SeFa. For example, when editing the size of the bird, our algorithm better preserves the background appearance. Our method also achieves better performance than SeFa from the perspective of smiling. It can also be seen that our method captures the same rotation and pose attributes as SeFa. The reason for this may be that ICA under orthogonal constraint and PCA can discover exactly the same most representative semantics (rotation for the bird model and pose for the face model).

We discovered that our algorithm is able to better remove the smile than SeFa, which is illustrated in Fig. 5.6. We can observe that our method can edit the smile on the face while preserving other attributes, such as pose, background and hair. However, when removing the smile using SeFa, background and hair style are changed.

**Quantitative analysis.** We perform quantitative analysis on two captured semantics, i.e., smile and size, in order to further validate the effectiveness of our algorithm. To quantitatively evaluate the direction about smile identified by our approach, following [94], we train a smiling classifier on the Multi-Modal CelebA-HQ data set with ResNet-50 network [37]. The result of the classifier may indicate whether the discovered direction is correlated with the desired attribute. Note that since many significant attributes, such as gender, age and glasses, are only manipulated by textual descriptions, we do not train the corresponding predictors. For the quantitative analysis of the direction about the size of the bird, inspired by [82], we use the cosine distance between the CLIP embeddings of the
Figure 5.4: Semantically-understandable semantics discovered in DTGAN that was trained on (1) the CUB bird (four top rows) and (2) Multi-Modal CelebA-HQ data sets (four bottom rows).
Figure 5.5: Qualitative comparison of the meaningful latent-space directions discovered by (a) SeFa [95] and (b) our proposed algorithm on (1) the CUB bird (four top rows) and (2) Multi-Modal CelebA-HQ (four bottom rows) data sets.
Figure 5.6: Qualitative comparison of the removing-smile latent-space direction discovered by (a) SeFa [95] and (b) our proposed algorithm on the Multi-Modal CelebA-HQ data set. –Smile indicates the removing-smile image.

Table 5.1: Quantitative analysis of the size and smiling semantics discovered by our method and SeFa [95]. The best scores are in bold.

<table>
<thead>
<tr>
<th>Method</th>
<th>size ↑</th>
<th>smile ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>SeFa</td>
<td>0.25</td>
<td>0.54</td>
</tr>
<tr>
<td>Ours</td>
<td>0.27</td>
<td>0.56</td>
</tr>
</tbody>
</table>

edited birds and the text ‘a large bird’ as the evaluation metric. The reason behind using the CLIP loss is that it can effectively guide image manipulation through a text prompt. The results are depicted in Table 5.1. We can see that our algorithm outperforms SeFa on both the smile and the size of the bird, which suggests its effectiveness. The above results demonstrate that based on the Good latent codes found by our well-trained classification model, we can adopt our presented algorithm to acquire a wealth of semantically-diverse and perceptually-realistic samples.

**Human evaluation.** We conduct a human test on the Multi-Modal CelebA-HQ data set to compare our method with SeFa. We randomly select 100 successful synthesized faces acquired by DiverGAN while employing the semantically-understandable directions (i.e., smile and hair) found by these two approaches to edit them. Users are asked to choose the sample with the most accurate change. Simultaneously, the final results are calculated by two judges for fairness. As illustrated in Fig. 5.7, our method performs better than SeFa with respect to the control of smile and hair, which demonstrates the superiority of our proposed algorithm.

5.4.5 Results of background flattening

To confirm the effectiveness of the proposed background-flattening loss (BFL), we apply it to optimize the meaningful latent-space directions identified in
Figure 5.7: Human test results (ratio of 1st) of SeFa [95] and our proposed method with respect to the smile and hair semantics on the Multi-Modal CelebA-HQ data set.

DiverGAN that was trained on the CUB bird data set. Fig. 5.8 provides two examples. By comparing the initial directions with the new semantics obtained by the BFL, we can see that the background is significantly improved and other image contents are maintained. As can be observed in the background-removal row, background flattening effectively removes the background while keeping the bird unchanged, which validates the effectiveness of the BFL. It indicates that the presented BFL can be employed for existing latent-code manipulation approaches to improve or remove the background of an edited sample.
Figure 5.8: Visualization of background flattening and background removal for the meaningful directions acquired by our proposed method.
5.5 CONCLUSION

In this chapter, we propose several techniques to better understand and exploit the latent space of a conditional text-to-image GAN model. More specifically, we introduce the independent component analysis (ICA) algorithm under an additional orthogonal constraint that can extract both independent and orthogonal components from the pretrained weight matrix of the generator as the semantically-interpretable latent-space directions. In addition, we mathematically analyze the correspondences between SeFa, GANSpace and regular PCA, since they utilize algorithmically simple techniques to derive useful semantics. Furthermore, we design a background-flattening loss (BFL) to optimize the background appearance in an edited sample. We evaluate our presented approaches on the recent single-stage text-to-image GAN models that were pre-trained on three benchmark data sets, i.e., the CUB bird, Multi-Modal CelebA-HQ and MSCOCO data sets. Extensive experimental results demonstrate that our introduced algorithm can derive meaningful semantic properties in the latent space of DTGAN, DF-GAN and DiverGAN. More importantly, our approach leads to a more powerful control over the latent codes than SeFa, which validates the effectiveness of our proposed method. Furthermore, we show that SeFa, GANSpace and regular PCA obtain almost the identical directions, supporting our claim in Section 5.3.2. Moreover, we discover that our BFL is able to effectively improve the background appearance in an edited sample while keeping other image contents unchanged, which confirms its effectiveness.