NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCES
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS

BSc THESIS

MVU-GAN: Unfolding the Latent Space of GANs

Pantelis K. Papageorgiou

Supervisor: **Yannis Panagakis,** Associate Professor

ATHENS

September 2021

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

MVU-GAN: Ξεδιπλώνοντας τον Λανθάνων Χώρο των GANs

Παντελής Κ. Παπαγεωργίου

Επιβλέπων: **Γιάννης Παναγάκης,** Αναπληρωτής Καθηγητής

ΑΘΗΝΑ

ΣΕΠΤΕΜΒΡΙΟΣ 2021

BSc THESIS

MVU-GAN: Unfolding the Latent Space of GANs

Pantelis K. Papageorgiou
S.N.: 1115201700115

SUPERVISOR: **Yannis Panagakis,** Associate Professor

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

MVU-GAN: Ξεδιπλώνοντας τον Λανθάνων Χώρο των GANs

Παντελής Κ. Παπαγεωργίου
Α.Μ.: 1115201700115

ΕΠΙΒΛΕΠΩΝ: **Γιάννης Παναγάκης,** Αναπληρωτής Καθηγητής

ABSTRACT

Generative Adversarial Networks (GANs) are deep-learning-based generative models that learn to map noise latent vectors to high-fidelity images. Recent work has shown that the input latent space can be decomposed to semantically meaningful directions. Moving towards these directions corresponds to human interpretable image transformations. For example, from high level aspects such as face shape and general hair style, to smaller scale facial features to color schemes and microstructures, everything can be controlled by moving in the corresponding GAN latent space direction.

In order to achieve image editing by identifying latent space directions, previous state-of-the-art methods either based on supervised approaches or leverage the Principal Components Analysis (PCA) algorithm. The former have a tremendous disadvantage for the range of directions that can be explored, as they rely on a human-annotated set of scores for each attribute. The latter tend to use the same method with minor modifications, resulting in similar experimental observations.

In this work, we approach the problem of discovering semantic directions in an unsupervised way, using semidefinite programming to perform non-linear dimensionality reduction of the internal representation of GANs. In particular, we examine the generation mechanism of GANs and further utilize the famous algorithm of Maximum Variance Unfolding, also known as Semidefinite Embedding, to identify semantically meaningful directions by decomposing the pre-trained weights. Furthermore, extensive experiments are conducted on the state-of-the-art GAN architectures, StyleGAN and StyleGANv2, for 7 different datasets.

To our knowledge, this is the first work to approach this problem from the perspective of semidefinite programming. While the computational cost can be high, the results clearly demonstrate its superiority in various experiments, while in others they can be compared with the results of the most recent supervised and unsupervised methods. Code is available at https://github.com/PanPapag/MVU-GAN.

SUBJECT AREA: Computer Vision

KEYWORDS: GAN, Image Editing, Semantic Directions, Latent Space, Semidefinite Programming

ΠΕΡΙΛΗΨΗ

Τα Παραγωγικά Αντιπαλικά Δίκτυα (ΠΑΔ) είναι παραγωγικά μοντέλα που βασίζονται στην βαθιά μάθηση και μαθαίνουν να απεικονίζουν ένα θόρυβο λανθάνοντος διανύσματος σε εικόνες υψηλής αξιοπιστίας. Πρόσφατα έργα έχουν δείξει ότι ο λανθάνων χώρος εισόδου μπορεί να αποσυντεθεί σε κατευθύνσεις σημασιολογικά ουσιαστικές. Η μετακίνηση προς αυτές τις κατευθύνσεις αντιστοιχεί σε ερμηνεύσιμες, από τον άνθρωπο, μετατροπές εικόνας. Για παράδειγμα, από πτυχές υψηλού επιπέδου, όπως το σχήμα του προσώπου και το γενικό στυλ των μαλλιών, μέχρι τα μικρότερα χαρακτηριστικά του προσώπου έως τα χρώματα και τις μικροδομές, όλα μπορούν να ελεγχθούν μετακινώντας στην αντίστοιχη κατεύθυνση του λανθάνοντος χώρου ΠΑΔ.

Προκειμένου να επιτευχθεί η επεξεργασία εικόνας με τον εντοπισμό κατευθύνσεων του λανθάνοντος χώρου, οι σύγχρονες μέθοδοι είτε βασίζονται σε εποπτευόμενες προσεγγίσεις είτε αξιοποιούν τον αλγόριθμο PCA. Οι πρώτες έχουν ένα τεράστιο μειονέκτημα σχετικά με το εύρος των κατευθύνσεων που μπορούν να διερευνηθούν, καθώς βασίζονται σε ένα σύνολο ανθρωπίνων σχολιασμένων βαθμολογιών για κάθε χαρακτηριστικό. Οι τελευταίες τείνουν να χρησιμοποιούν την ίδια μέθοδο με μικρές τροποποιήσεις, με αποτέλεσμα παρόμοιες πειραματικές παρατηρήσεις.

Σε αυτήν την εργασία, προσεγγίζουμε το πρόβλημα της ανακάλυψης σημασιολογικών κατευθύνσεων χωρίς εποπτεία, χρησιμοποιώντας ημιπεριορισμένο προγραμματισμό για την εκτέλεση μη γραμμικής μείωσης διαστάσεων της εσωτερικής αναπαράστασης των ΠΑΔ. Συγκεκριμένα, εξετάζουμε τον μηχανισμό παραγωγής των ΠΑΔ και χρησιμοποιούμε περαιτέρω τον περίφημο αλγόριθμο Αναδίπλωσης Μέγιστης Διακύμανσης, επίσης γνωστό ως Ημιπεριορισμένη Ενσωμάτωση, για να εντοπίσουμε σημασιολογικά σημαντικές κατευθύνσεις αποσυνθέτοντας τα προεκπαιδευμένα βάρη. Επιπλέον, διεξάγονται εκτεταμένα πειράματα με τις πιο σύγχρονες αρχιτεκτονικές ΠΑΔ, StyleGAN και StyleGANv2, για 7 διαφορετικά σύνολα δεδομένων.

Από όσο γνωρίζουμε, αυτή είναι η πρώτη εργασία που προσεγγίζει αυτό το πρόβλημα από την οπτική του ημιπεριορισμένου προγραμματισμού. Ενώ το υπολογιστικό κόστος μπορεί να είναι υψηλό, τα αποτελέσματα αποδεικνύουν σαφώς την υπεροχή του σε διάφορα πειράματα, ενώ σε άλλα μπορούν να συγκριθούν με τα αποτελέσματα των πιο πρόσφατων εποπτευόμενων και μη εποπτευόμενων μεθόδων. Ο κώδικας είναι διαθέσιμος στο https://github.com/PanPapag/MVU-GAN.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Μηχανική Όραση

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: ΠΑΔ, Επεξεργασία Εικόνας, Σημασιολογικές Κατευθύνσεις, Λανθάνων Χώρος, Ημεπεριορισμένος Προγραμματισμός

ACKNOWLEDGEMENTS

I would like to thank my supervisor Professor Yiannis Panagakis for the valuable guidance and constant support towards the composition of this thesis. Also, I would like to thank Applied Memetic Ltd. which provided us with free access to AWS resources.

Finally, I would like to thank my family for their encouragement and support all through my studies.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

This thesis was written as a part of the BSc program of studies at the Department of Informatics and Telecommunications of the National and Kapodistrian University of Athens.

1. INTRODUCTION

The quality of images synthesized by deep generative models has improved dramatically over the past few years. Nowadays, Generative Adversarial Networks (GANs) [10] have achieved incomparable success in image synthesis and are getting widely adopted by digital artists. The state-of-the-art GANs, such as BigGAN [5] and StyleGAN [20, 21, 19], are able to produce high-fidelity synthetic images that can be indistinguishable from real ones. The rationale behind GANs is to utilize the adversarial method of training in order to learn a non-linear mapping from the input noise latent vectors to the distribution of the training data. After learning that mapping, GANs can be fed with randomly sampled noise vectors and produce images of high quality.

However, in real-world applications, GAN models are typically treated as black-boxes without a deeper understanding of the generative process taking place inside them. Several works have shown that the latent space of GANs present a useful vector arithmetic property, e.g. adding the latent codes of two images [20] or adding a learned vector to the input latent noise [25], results in a semantically modified image. Although, it is still not clear how the latent space is organized in terms of interpretability, these prior works motivate the researchers to understand the latent space of GANs.

A continuous active research on GANs interpretability aim to discover the underlying semantics of their latent spaces. A recent work [3] has shown that the intermediate neurons of a well-trained GAN are responsible for several parts of the visual world in a synthesized image, such as trees and doors for outdoor scenes generation. At the moment, the most recent works try discover interpretable directions in the latent space of GANs mainly following one of the two approaches: supervised or unsupervised methods.

The existing supervised approaches tend to first randomly sample a great amount of latent codes, then synthesize the corresponding set of images and human-annotate them with some predetermined labels, and finally train a linear classifier to learn a hyperplane on the latent space. In order to label the images for training, they either utilize pretrained attribute predictors [31, 9, 27] or exploit statistical information of the images, such as objects position and color hue. Although the supervised approaches may offer a high degree of control over the discovered directions, they also present severe limitations in terms of applicability. Specifically, a well annotated set of attributes is expensive to collect and can lead to non-deterministic results, e.g. sampling a different collection of latent codes may cause the classifier to learn different separating boundaries.

On the other hand, recently works [29, 14, 28] were published that follow the unsupervised approach to discover interpretable directions in the latent space of GANs. However, they either require model training [29] or data sampling [14]. The most recent one, SeFa [28] does not require neither of them, but it leverages the PCA algorithm, same as the GANSpace [14], in order to discover the desired latent semantics. Therefore, the aforementioned approaches tend to present similar results.

In this thesis, we propose a novel unsupervised algorithm that performs non-linear dimensionality reduction of the pre-trained weights of a generator, aiming at interpreting the latent semantics of GANs beyond traditional methods. We call it MVU-GAN, as the short of Maximum Variance Unfolding GAN. Similar to SeFa [28], our framework is independent of data sampling and model training. Instead of focusing on PCA to decompose the pretrained weights of a GAN model, MVU-GAN takes a step further into the relation of the modes of variability in the internal representation of GANs. In practice, GANs project an

input latent noise to a high quality image step by a step, and to be more precise layer by layer. As it is known from [20, 21] the mapping from one space to another is non-linear. While PCA may work poorly if the most important modes of variability is non-linear, Maximum Variance Unfolding (MVU) [30] tries to improve it using semidefinite programming (SDP). Our approach discovered some interesting and unexpected results that can clearly challenge the state-of-the-art unsupervised methods, as well as demonstrate superiority compared to the range of the latent semantics learned by supervised algorithms. Some results of our method are shown in 1.1

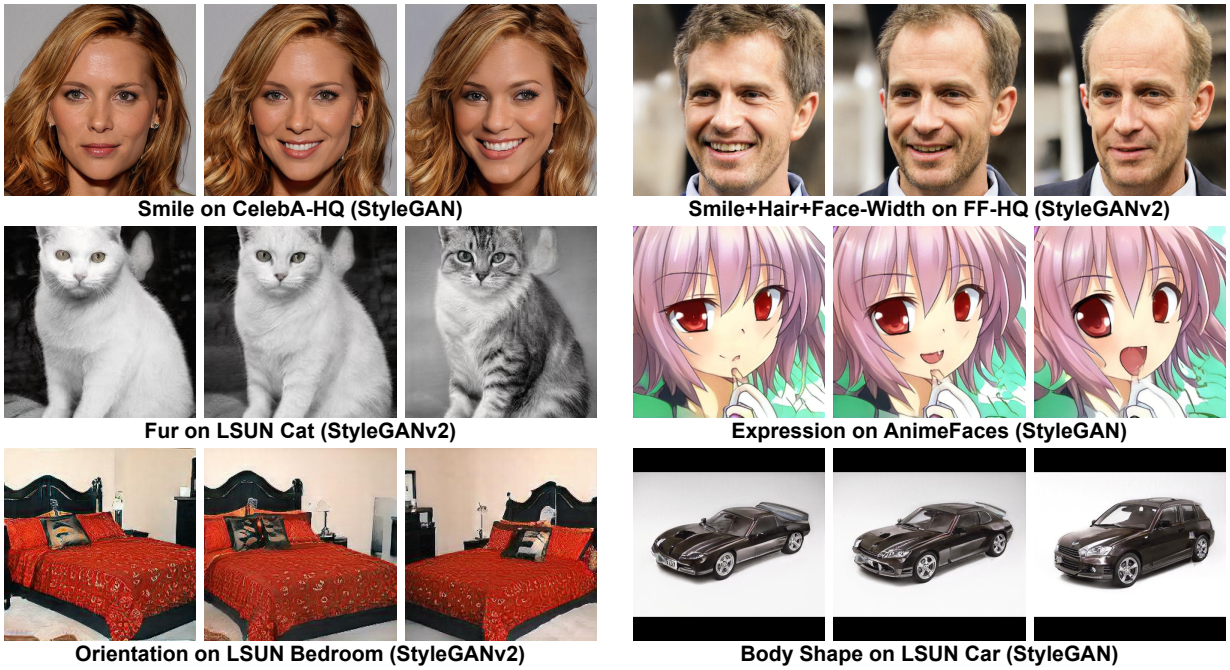

Figure 1.1: A sample of semantically meaningful directions unsupervisedly discovered by our method MVU-GAN. For each set of images, the middle image is the original one while the left and the right are the synthesized images produced by moving the latent code backwards or towards the explored boundary.

Our contributions are summarized as follows:

- We propose a novel unsupervised method that leverages semidefinite programming to explore semantically meaningful directions in the latent space of a pre-trained GAN.
- We show that our algorithm is able to discover different variations of a certain interpretable direction, leading to results that appear for the first time in the literature.
- We show that the learned directions are well-disentangled, comparing our approach with the state-of-the-art supervised and unsupervised method.
- We conduct extensive experiments on the state-of-the-art style-based generator models trained on diverse datasets, proving the effectiveness of our algorithm.

2. RELATED WORK

In this section, we describe the relevant research areas and explain the scientific context of our study.

2.1 Generative Adversarial Networks

GAN [10] has become the state-of-the-art deep generative model paradigm in producing photo-realistic images [26, 1, 4, 18, 20, 21, 19, 5, 34]. GAN consists of two networks. The first one is the generator, which maps the input latent noise to the complex data distribution of the training dataset. The second one is the discriminator, which aims at distinguishing the synthetic data from the real one. In fact, GAN is a competing game between two players (neural networks) played in adversary in order to reach Nash Equilibrium. In recent years the progress of GANs has grown dramatically from different perspectives, e.g. by improving the training process [26], or the discriminator [8], or by carefully modifying the generator's architecture [18, 33], or by choosing a more descriptive loss function [1, 17]. The state-of-the-art models, such as StyleGAN [20], StyleGANv2 [21], StyleGANv3 [19] and BigGAN [5] are able to produce synthetic images of high-quality and high-fidelity, usually identical to those of the real world.

Among them, the StyleGAN architectures, in which we focus on this work, introduced a new way to feed the latent code into the generator. All previous models composed of a deep convolutional neural network generator where the input latent noise was fed into the first convolution layer through an affine transformation [25, 18, 1]. This approach was recently improved by the style-based generator [20, 21, 19]. As shown in Figure 2.1, the input latent code $\mathbf{z} \in \mathcal{Z}$ is first mapped to an intermediate latent space $\mathcal{W}$, and then fed into each convolution block through Adaptive Instance Normalization (AdaIN) [13] operation. It has been proved that this mapping can "unwrap" $\mathcal{W}$, so that the factors of variation become more linear. Consequently, in addition to synthesizing photo-realistic images, the generator is also able to organize a less entangled latent space, resulting in easier and more accurate semantic editing in the latent space.

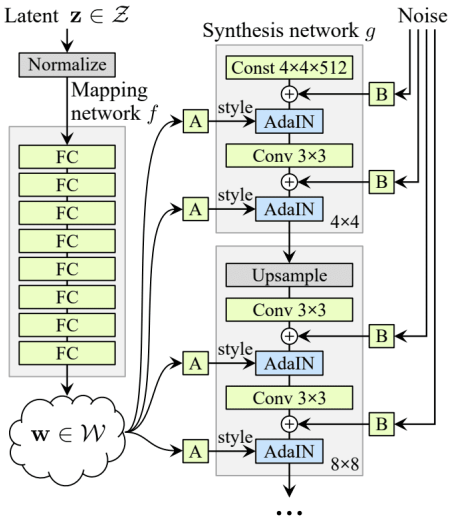

Figure 2.1: Style-based generator of StyleGAN [20]

2.2 Semantic Directions in the Latent Space

Ever since the first GAN models appeared, it has been shown that the latent space of GANs presents interesting vector arithmetic properties, allowing for straightforward manipulation of semantic qualities of the generated samples [25]. Since moving towards these interpretable directions in the latent space would enable an effortless way to perform effective image editing, the discovery of such directions currently motivates the researchers to go beyond the limits of the state-of-the-art methods in the field.

Prior work, such as InfoGAN [7], proposed the addition of regularizers into the training process in order to learn explicitly interpretable factorized vectors. It has recently been found that the GANs encode semantically meaningful representations in the intermediate feature space [3], as well as in the input latent space [15, 29, 27, 14, 28, 24]. Shen *et al.* [27] propose a supervised framework which requires selecting a collection of images and annotate them with human-defined labels to train a classifier. The major drawback of supervised approaches is their need of human annotators or pre-trained models, which can be time-consuming and expensive. More importantly, a supervised method can only lead to the discovery of directions which researchers ultimately expect to identify, e.g. age, pose, glasses etc. for semantic face editing. Jahanian *et al.* [15] and Plumerault *et al.* [24] developed self-supervised approaches, but they are limited to discovering affine transformations that correspond to simple image augmentations such as rotation and zooming.

On the other hand, exploring the latent space of GANs in an unsupervised manner has been proven to be as effective as supervised approaches, if not more so, while not suffering from the aforementioned weaknesses they present. Voynovand and Babenko [29] jointly optimize a matrix A and a reconstructor R, so that the columns of A will correspond to the discovered directions, while the reconstructor's purpose is to reproduce the shift in the latent space that causes a given image transformation. Härkönen *et al.* randomly sample a set of latent vectors $\mathbf{z}_{1:N}$, compute the corresponding $\mathbf{w}_i = M(\mathbf{z}_i)$, and then perform PCA, to find semantic directions in the latent space. However, both of these methods either require model training or data sampling. The most recent work proposed by Shen and Zhou [28] shows that the pre-trained weights of a GAN model, denoted by $\mathbf{A}$, can be factorized using PCA so that the eigenvectors of $\mathbf{A}^T\mathbf{A}$ will correspond to the explored directions. Although, this algorithm requires neither training models nor data sampling, it is still approaches the problem of identifying interpretable directions in the latent space of GANs on a linear fashion. Specifically, it tries to perform linear dimensionality reduction on high-dimensional vectorial input data, while the latent space of GANs is generally treated as Riemannian manifold [2, 6, 22].

Differently, we study the generation process of GANs and propose a method, independent of model training and data sampling, that leverages the MVU algorithm to perform non-linear dimensionality reduction of the internal representation of GANs. This approach leads to accurately decomposing the most important modes of variability and exploring semantically meaningful directions in a completely unsupervised way.

3. FRAMEWORK OF MVU-GAN

In this section, we introduce the framework of MVU-GAN. Firstly, we are going to dive into the generation mechanism of GANs and examine how to perform image editing by manipulating the latent space of a well trained GAN. Finally, a thorough analysis of MVU algorithm will be presented, as well as its use to decompose the model weights. As this work focuses on StyleGAN architectures, detailed references will be made in each subsection.

3.1 Background

3.1.1 Generation Process of GANs

The generator $G(\cdot)$ learns a non-linear mapping from the d-dimensional input latent space $\mathbf{z} \in \mathcal{Z}$, where $\mathcal{Z} \subseteq \mathbb{R}^d$, to the output higher dimensional image space $\mathcal{I} \subseteq \mathbb{R}^{H \times W \times C}$. The aforementioned description can be mathematically formulated as $G(\mathbf{z}) = I$, where I constitutes the output image. In order to model the generator as a non-linear function, the state-of-the-art GANs [25, 18, 5, 20, 21, 19] have adopted convolutional neural networks (CNNs) as its architecture. A GAN generator consists of multiple convolutional layers, each of which learns a non-linear transformation from one space to another. In particular, the generator can be decomposed into L intermediate layers $G_1, G_2, \dots G_L$. The first layer is fed with the input latent noise $\mathbf{z}$ and produces a feature map $y_1 = G_1(\mathbf{z})$. The remaining layers take as input the output of the previous one, producing a new feature map $\mathbf{y_i} = \hat{G_1}(\mathbf{z}) \equiv G_i(\mathbf{y_{i-1}})$. The last layer generates the final image $I = G_L(\mathbf{y_{L-1}})$. In a StyleGAN model [20, 21], the authors modify the generator's input so that it no longer takes a point from the latent space. Instead, the model has a learned constant tensor $T \subseteq \mathbb{R}^{4 \times 4 \times 512}$ to start the image synthesis process. Moreover, given a latent code $\mathbf{z} \in \mathcal{Z}$, a non-linear mapping network $M : \mathcal{Z} \to \mathcal{W}$, where $M(\cdot)$ is an 8-layer multilayer (MLP), first produces $\mathbf{w} \in \mathcal{W}$. The so-called style vector $\mathbf{w}$ is then transformed and incorporated into each block of the generator after the convolutional layer via AdaIN [13] operation:

$$\mathbf{y_i} = G_i(\mathbf{y_{i-1}}, \mathbf{w}), \quad \text{with } \mathbf{w} = M(\mathbf{z}) \tag{3.1}$$

The Eq. (3.1) can be further formulated as an affine transformation:

$$G_i(\mathbf{y_{i-1}}, \mathbf{w}) = \mathbf{y_i} \equiv \mathbf{A_i w} + \mathbf{b_i}, \tag{3.2}$$

where $\mathbf{y_i} \in \mathbb{R}^m$, the m-dimensional output of the i_{th} layer. $\mathbf{A_i} \in \mathbb{R}^{m \times d}$ and $\mathbf{b_i} \in \mathbb{R}^m$ denote the weight and bias used in the transformation of the i_{th} step $G_i(\cdot)$ of the generator.

3.1.2 Manipulation in the Latent Space

Recent work has shown that the latent space of GANs encodes information semantically meaningful [9, 15, 27, 14, 29]. By exploiting the vector arithmetic property [25], these semantic directions can be applied to image editing. To this end, our goal is to identify the direction $\theta \in \mathbb{R}^d$ that corresponds to an interpretable image transformation in the latent space of a pre-trained GAN, so that the manipulation can be achieved via the following formula:

$$I = G(\mathbf{z} + \alpha\theta), \tag{3.3}$$

where $\alpha \in \mathbb{U}[-b, b]$ is a scalar that controls the degree of change of the target direction.

In the StyleGAN model, the manipulation of the latent space can be performed as:

$$\begin{aligned} \mathbf{w} &= G_{map}(\mathbf{z}), \\ I &= G_{syn}(\mathbf{w} + \alpha\theta) \end{aligned} \tag{3.4}$$

In this thesis, we focus on the latent space $\mathcal{W}$ of StyleGAN, so we are going to use $G_{syn}(\cdot)$ and $G(\cdot)$ interchangeably.

3.2 Unfolding the Latent Space

3.2.1 Motivation

Shen and Zhou [28] proved that the manipulation process is instance independent. Specifically, let us take the affine transformation of Eq. (3.2) and semantically edit the latent space as shown in Eq. (3.3) and (3.4):

$$\begin{aligned} \mathbf{y_i}' &= G_i(\mathbf{w} + \alpha\theta) \\ &= \mathbf{A_i}\mathbf{w} + \mathbf{b_i} + \alpha\mathbf{A_i}\theta \\ &= \mathbf{y_i} + \alpha\mathbf{A_i}\theta \end{aligned} \tag{3.5}$$

We observe from Eq. (3.5) that the desired image editing can be achieved only by adding the term $\alpha\mathbf{A_i}\theta$ in a arbitrary given step i of the generator $G(\cdot)$. Subsequently, as mentioned in their work [28], the weight parameter $\mathbf{A_i}$ should contain the essential knowledge of the image variation. SeFa [28] is trying to discover the most important directions by solving the following optimization problem:

$$\theta^* = \underset{\theta \in \mathbb{R}^d:\, \theta^T\theta = 1}{\arg\max} \; ||\mathbf{A_i}\theta|| \tag{3.6}$$

The intuition is to explore those directions which will cause large variation after the projection of $\mathbf{A_i}$. Finally, the solutions to the optimization problem of Eq. (3.6) that correspond to the k most important directions are the k largest eigenvectors of $\mathbf{A_i}^T\mathbf{A_i}$. In fact, this method utilizes the PCA algorithm to compute the similarities between the high dimensional data inputs $\mathbf{X}$ using the linear kernel $\mathbf{X}^T\mathbf{X}$, where $\mathbf{X} \equiv \mathbf{A_i}$, and then perform eigenvalue decomposition on that linear kernel matrix.

3.2.2 Maximum Variance Unfolding (MVU)

Maximum Variance Unfolding (MVU) can be viewed as a non-linear generalization of PCA, which tries to improve its performance if the most important modes of variability in the input data are non-linear. Instead of assuming the existence of a kernel to start, MVU defines an optimization problem that tries to learn the optimal kernel matrix. An optimal kernel is considered to be the one that captures the similarity in local patch, while preserving the geometry of the space. In practice, MVU is structured as a semidefinite programming problem that aims to find the Gramian matrix K that maximizes the pairwise distance of every pair of points, while preserving the distances of neighbor points. Intuitively, as the name implies, it tries to pull the points of the high-dimensional input data as far away from each other as possible and therefore “unfold” the manifold.

3.2.2.1 Optimization Formula

The aforementioned objective and parameters can be defined by the following convex optimization problem:

$$\begin{aligned} &\max_K \sum_i^n \sum_j^n ||Y_i - Y_j||^2 \text{ such that:} \\ &1.\ ||Y_i - Y_j||^2 = ||X_i - X_j||^2 \text{ for all } (i,j) \text{ with } \eta_{ij} = 1 \\ &2.\ |\sum_i^n Y_i|^2 = 0 \\ &3.\ K \succeq 0, \end{aligned} \tag{3.7}$$

where X is the original high-dimensional data and Y is their respective representation in a lower dimensional space.

However, in order to be able to apply semidefinite programming to learn the optimal kernel matrix K, MVU redefines the problem in terms of Gramian matrices for the two spaces. It aims to find the Gramian matrix K that maximizes the distances between all data points except those that are nearest to each other. In order to discover the points nearest to one another, the edge matrix E is first constructed using the K-Nearest Neighbor (KNN) algorithm. The edge matrix E, otherwise known as neighborhood graph, has the following form:

$$E = \begin{bmatrix} \eta_{11} & \dots & \eta_{1n} \\ \vdots & \ddots & \vdots \\ \eta_{n1} & \dots & \eta_{nn} \end{bmatrix}, \eta_{ij} \in \{0,1\}, \sum_j^n \eta_{ij} = k$$

Each entry of the edge matrix E gets the value 1 if the corresponding pair of points are neighbors, otherwise it gets the value 0. The number of neighbors k determines the degree of connectivity of the graph. It has been shown that if the data is sampled well enough by selecting the optimal value for the hyperparameter k, the resulting neighborhood graph E is a discrete approximation of the underlying manifold. Moreover, it is important to know if the graph E is connected or not. In case of a disconnected neighborhood graph, the distance between the disconnected points will go towards infinity, while trying to optimize the objective function by maximizing the distance between the points. This edge case can be handled by checking the sign of the eigenvalues of the Laplacian of E, ensuring that all of them are positive.

Now, let G, K be the Gramian matrices of X and Y, such that $G_{ij} = X_i X_j$ and $K_{ij} = Y_i Y_j$. Having defined the local isometry property by the entries of E with value of 1, the first constraint of (3.7) can be rewritten in terms of the Gramian matrices as follows:

$$\begin{aligned} ||X_i - X_j||^2 &= ||Y_i - Y_j||^2 \\ X_i^T X_i + X_j^T X_j - X_i^T X_j - X_j^T X_i &= Y_i^T Y_i + Y_j^T Y_j - Y_i^T Y_j - Y_j^T Y_i \\ G_{ii} + G_{jj} - G_{ij} - G_{ji} &= K_{ii} + K_{jj} - K_{ij} - K_{ji} \\ G_{ii} + G_{jj} - 2G_{ij} &= K_{ii} + K_{jj} - 2K_{ij} \\ &\Rightarrow \\ \eta_{ij}(G_{ii} + G_{jj} - 2G_{ij}) &= \eta_{ij}(K_{ii} + K_{jj} - 2K_{ij}) \end{aligned} \tag{3.8}$$

In addition to this, the second constraint of (3.7) demands Y to be centered at the origin:

$$\begin{aligned} |\sum_{i}^{n} Y_i|^2 = 0 &\Leftrightarrow (\sum_{i}^{n} Y_i)^T (\sum_{i}^{n} Y_i) = 0 \\ &\Leftrightarrow (\sum_{i}^{n} Y_i)^T (\sum_{j}^{n} Y_j) = 0 \\ &\Leftrightarrow \sum_{i}^{n}\sum_{j}^{n} Y_i^T Y_i = 0 \\ &\Leftrightarrow \sum_{i}^{n}\sum_{j}^{n} K_{ij} = 0 \end{aligned} \tag{3.9}$$

Thus, centering the data, as shown by (3.9), forces the sum of all the points in the kernel matrix K to be equal to zero.

The objective function of (3.7) can be rewritten completely in the form of the Gramian matrix:

$$\begin{aligned} \sum_{i}^{n}\sum_{j}^{n} ||Y_i - Y_j||^2 &= \sum_{i}^{n}\sum_{j}^{n} (K_{ii} + K_{jj} - 2K_{ij}) \\ &= \sum_{i}^{n}\sum_{j}^{n} (K_{ii} + K_{jj}) \\ &= \sum_{i}^{n}\sum_{j}^{n} K_{ii} + \sum_{i}^{n}\sum_{j}^{n} K_{jj} \\ &= n\sum_{i}^{n} K_{ii} + n\sum_{j}^{n} K_{jj} \\ &= 2nTr(K) \\ &\Rightarrow \\ \max(\sum_{i}^{n}\sum_{j}^{n} ||Y_i - Y_j||^2) &= \max(2nTr(K)) \\ &= \max(Tr(K)) \end{aligned} \tag{3.10}$$

The result from (3.10) shows that maximizing the distance of all points not connected in the neighborhood graph E is equivalent of maximizing the Trace of the Gramian matrix K.

Finally, the optimization problem of (3.7) can be formulated as the following semidefinite program:

$$\begin{aligned} &\max_{K} \quad Tr(K) \\ &\text{subject to} \quad K \succeq 0, \sum_{i}^{n}\sum_{j}^{n} K_{ij} = 0 \\ &\text{and} \quad \eta_{ij}(G_{ii} + G_{jj} - 2G_{ij}) = \eta_{ij}(K_{ii} + K_{jj} - 2K_{ij}) \end{aligned} \tag{3.11}$$

3.2.2.2 Algorithm

To sum up, MVU learns an optimal kernel matrix K that provides a mapping from high-dimensional input space to a lower-dimensional output space in the following steps:

1. The neighborhood graph E is constructed using the K-NN algorithm. To ensure that E is not disconnected, the eigenvalues of the Laplacian of E must all be positive.
2. Semidefinite programming is applied to “unfold” the neighborhood graph.
3. After the Gramian matrix K is learned by solving the semidefinite program defined in (3.11), the output space Y can be obtained via Cholesky decomposition.

3.2.3 Implementation on StyleGAN Models

Our procedure can be simply applied on StyleGAN models [20, 21]. We have already seen in Sec. 3.2.1 that the matrix A_i contains the necessary knowledge of the image variation in order to achieve effective image editing. For the target layers we intend to decompose, we concatenate their weight matrices (i.e. A_i from Eq. (3.2)) along the first axis. Motivated by the method of SeFa [28], we improve it by suggesting that instead of using the linear kernel $X^T X$, where $X \equiv A_i$, it is more effective to learn the optimal kernel matrix K by using the MVU algorithm described in Sec. 3.2.2. Finally, having learned the Gramian matrix K, the top-k most important directions can be obtained by choosing the k largest eigenvectors of K.

4. EXPERIMENTS

In this section we evaluate the proposed MVU-GAN in terms of both qualitative and quantitative results. Our experiments are performed on style-based generators and aim to discover a rich set of interpretable directions on a wide range of datasets. We compare MVU-GAN with the existing state-of-the-art supervised and unsupervised approaches and demonstrate its effectiveness through evaluation by a complete set of metrics.

4.1 Results on Diverse Datasets and Generator Models

4.1.1 Datasets and Models

We conduct our experiment on the state-of-the-art style-based generators, StyleGAN [20] and StyleGANv2 [21], trained on seven common datasets:

1. Flickr-Faces-HQ (FFHQ) [20], containing 1024×1024 images with significant variation in terms of age, ethnicity and background colors, as well as micro-structures such as eyeglasses, hats, sunglasses etc. Here, we use both StyleGAN and StyleGANv2 available at karras2019stylegan-ffhq-1024x1024.pkl and stylegan2-ffhq-config-f.pkl respectively.
2. Large-scale CelebFaces Attributes (CelebA-HQ) dataset [23], containing 1024×1024 celebrity images, each with 40 attribute annotations. We use StyleGAN available online at karras2019stylegan-celebahq-1024x1024.pkl.
3. AnimeFaces dataset [16] at 512×512 resolution. We use StyleGAN available online at stylegan_animefacee512.pth.
4. StyleGAN and StyleGANv2 trained with LSUN Car [32] dataset at 512×384 resolution. The generator models can be found online at karras2019stylegan-cars-512x384.pkl and stylegan2-car-config-f.pkl.
5. StyleGAN and StyleGANv2 trained with LSUN Cat [32] dataset at 256×256 resolution. The generator models can be found online at karras2019stylegan-cats-256x256.pkl and stylegan2-cat-config-f.pkl.
6. LSUN Bedroom [32] dataset, containing 256×256 images. Here, we use StyleGAN available at karras2019stylegan-bedrooms-256x256.pkl.
7. LSUN Church [32] dataset, containing 256×256 images. Here, we use StyleGANv2 available at stylegan2-church-config-f.pkl.

4.1.2 Results on StyleGAN Models

As described in Sec.3.2.3 our method can decompose either a single layer or multiple layers concatenated along the first axis. Our experiments justify the observation by [20] that style-based generators tend to learn image construction in a hierarchical manner. In particular, Fig. 4.1 shows that bottom layers ($4^2 - 8^2$ resolution) control high-level aspects of the image such as pose, camera viewpoint, general hairstyle etc. Moreover, the middle

Figure 4.1: Hierarchical semantically meaningful directions discovered by MVU-GAN in the style-based generator of StyleGAN [20] and StyleGANv2 [21].

layers ($16^2 - 32^2$ resolution) are responsible for structural features, while the top layers ($64^2 - 1024^2$ resolution) handle the color schemes and microstructures. Taking cats as an example, bottom layers control the camera's viewpoint, middle layers determines the cat's expression, while the top layers handle the color of the fur.

4.1.3 Multiple Variants

As we have discussed in Sec.3.2.2.1, the MVU algorithm is sensitive to hyperparameter k. An optimal choice of hyperparameter k can lead to the neighborhood graph being a discrete approximation of the underlying manifold. We performed experiments with different values for k and observed that MVU-GAN is able to discover multiple variants of a particular attribute when it decomposes a given set of layers. In Fig.4.2 we show that for StyleGAN model trained on (a) LSUN-Cat and (b) LSUN-Car datasets, MVU-GAN is able to interpret two different variants of expression and shape respectively.

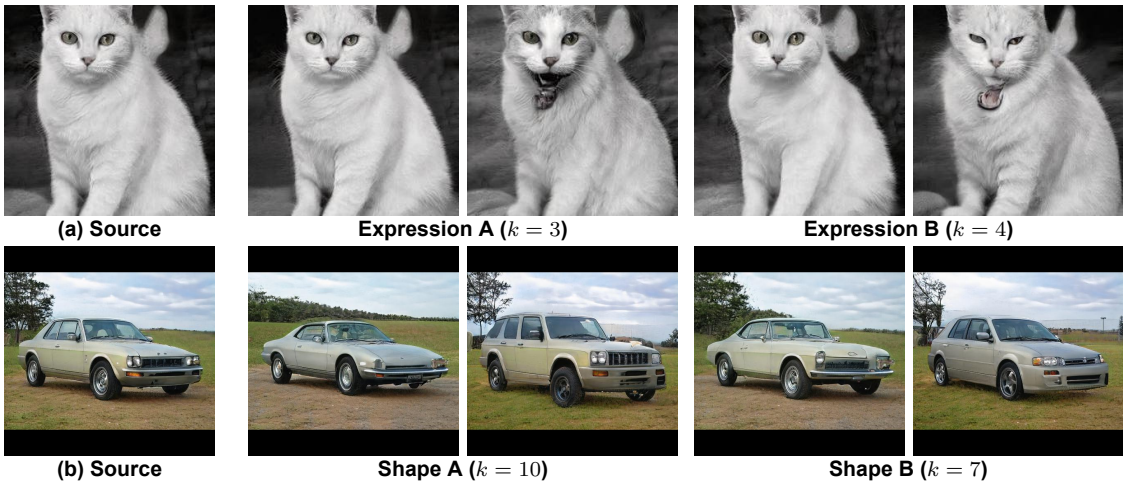

Figure 4.2: Multiple variants of discovered directions found by MVU-GAN for different values of hyperparameter k.

4.2 Comparison with Supervised Approach

We compare the proposed framework of MVU-GAN with the state-of-the-art supervised method, InterFaceGAN [27]. In their work, Shen et al. [27] mention that the effectiveness of InterFaceGAN is based on the assumption that for any binary attribute, there exists a hyperplane in the latent space of a well trained generator that can separate the latent space data points into two groups. The samples of one group will have the given attribute, while the others will not. For this reason, we choose to conduct experiments on face generation models due to well definition of facial attributes. In particular, we make comparison between MVU-GAN and InterFaceGAN on CelebA-HQ and FF-HQ datasets using StyleGAN's style-based generator.

4.2.1 Qualitative Results

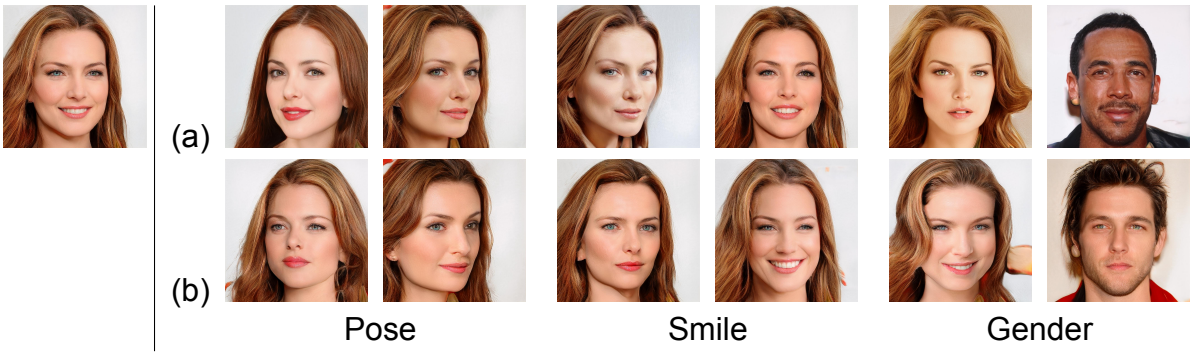

Figure 4.3: Qualitative comparison of the latent semantics found by (a) the supervised method, InterFaceGAN [27] and (b) our proposed framework MVU-GAN from the StyleGAN model [20] trained on CelebA-HQ [23] dataset.

In Fig.4.3 and 4.4, we visualize some manipulation results by moving towards certain discovered directions. Our approach, although completely unsupervised, seems to be superior to InterFaceGAN in a wide range of experiments. For instance, we can tell that the gender manipulation of InterFaceGAN on CelebA-HQ dataset (Fig.4.3) clearly does not highlight the expected results (a blond white girl is transformed into a black man),

Figure 4.4: Qualitative comparison of the latent semantics found by (a) the supervised method, InterFaceGAN [27] and (b) our proposed framework MVU-GAN from the StyleGAN model [20] trained on FF-HQ [20] dataset.

while our approach seems to be way more invariant in the other aspects of the image, e.g. race, pose, image color, etc. Moreover, our experiments from the StyleGAN model trained on FF-HQ dataset (Fig.4.4) demonstrate similar findings. We observe that InterFaceGAN has clearly explored entangled directions, as changing a particular identified semantic also changes other distinctive features. In contrast, our framework MVU-GAN discovers disentangled semantics, leading to accurate attribute manipulation.

4.2.2 Quantitative Results

4.2.2.1 Re-scoring Analysis

Re-scoring analysis, proposed by [28], aims to quantify the degree of change of a particular attribute, caused by the manipulation of the corresponding identified direction. We follow the process described in [28] and train an attribute predictor on CelebA-HQ [23] dataset with ResNet-50 [11] structure, as in [27]. First, we randomly sample $2K$ images and manipulate them along a given discovered direction. Then, using the pretrained attribute predictor we qualitatively evaluate the semantic score of each feature in such a manipulation process.

Table 4.1: Quantitative comparison using re-scoring analysis between the state-of-the-art supervised approach InterFaceGAN [27] and our proposed method MVU-GAN for the StyleGAN model [20] trained on CelebA-HQ dataset. Each row quantifies the change of the semantic score after moving into a certain direction in the latent space.

	Pose	Gender	Age	Eyeglasses	Smile
Pose	0.51	0.05	-0.09	0.00	0.10
Gender	-0.01	0.55	0.26	0.20	-0.04
Age	-0.02	0.39	0.50	0.20	0.12
Eyeglasses	-0.03	0.41	0.24	0.27	-0.01
Smile	-0.01	-0.10	0.01	-0.04	0.58

(a) InterFaceGAN

	Pose	Gender	Age	Eyeglasses	Smile
Pose	0.55	0.06	0.12	0.09	0.16
Gender	-0.01	0.65	0.02	-0.03	-0.01
Age	-0.02	0.03	0.45	0.15	0.01
Eyeglasses	0.05	0.19	0.09	0.49	0.03
Smile	0.03	0.01	-0.04	0.00	0.75

(b) MVU-GAN

Tab.4.1 shows the results where we have the following key observations. MVU-GAN seems to not be able to discover a well-disentangled direction corresponding to pose attribute in contrast to InterFaceGAN. However, our proposed method is clearly more robust in every other feature manipulation, e.g. gender, age, eyeglasses and smile. The reason is that MVU-GAN can optimally "unfold" the pretrained weights of the generator when a good value of hyperparameter k is selected.

4.2.2.2 Fréchet Inception Distance

Fréchet Inception Distance (FID) was proposed by Heusel et al. [12] as an improvement of the existing Inception Score (IS). In practice, FID is a metric for evaluating the quality of synthetic images and has proven to be effective in measuring GANs' performance. In our work, we are going to use FID score to evaluate our results in the following simple way. First, we randomly sample $50K$ synthetic images and manipulate them towards the discovered directions (pose, gender, age, eyeglasses, smile). As we have defined in Sec.3.1.2, the degree of change of a target direction is controlled by the parameter α. For each image manipulation we randomly select with a probability of 0.5 the parameter α to be equal either $-\beta$ or β. We thus end up creating a new dataset consisting of images that have been manipulated at the maximum degree of change towards the given directions. Then, we use $50K$ images drawn randomly from the training set to finally compute the FID score between the two datasets. Tab. 4.2 summarizes our results. We can tell that our proposed framework MVU-GAN outperforms the existing state-of-the-art supervised method InterFaceGAN in terms of image quality evaluated accordingly the FID score.

Table 4.2: Quantitative comparison using FID score between the state-of-the-art supervised approach InterFaceGAN [27] and our proposed method MVU-GAN for the StyleGAN model [20] trained on (a) CelebA-HQ dataset and (b) on FF-HQ dataset.

Method	FID
InterFaceGAN	6.25
MVU-GAN	**6.10**

(a)

Method	FID
InterFaceGAN	5.81
MVU-GAN	**5.52**

(b)

4.2.3 Correlation between Attributes

Karras et al. [20] proposed two metrics, Perceptual Path Length (PPL) and Linear Separability, in order to measure the disentanglement of the latent space. In this thesis, we focus more on studying the relationship between explored semantically meaningful directions and how they interact with each other. To do so, we propose a simple metric, firstly described in [27]. Given two identified directions θ_1 and θ_2, where θ_1, θ_2 stand for unit vectors, we compute the cosine similarity as $\cos(\theta_1, \theta_2) = \theta_1^T \theta_2$.

Table 4.3: Correlation matrices of attribute directions discovered using supervised method InterFaceGAN.

	Pose	Gender	Age	Eyeglasses	Smile
Pose	1.00	0.02	0.05	0.08	-0.13
Gender	-	1.00	0.26	0.64	-0.16
Age	-	-	1.00	0.31	0.12
Eyeglasses	-	-	-	1.00	0.13
Smile	-	-	-	-	1.00

(a) Correlation matrix of attribute boundaries from the StyleGAN model trained on CelebA-HQ dataset.

	Pose	Gender	Age	Eyeglasses	Smile
Pose	1.00	0.00	0.03	-0.08	-0.03
Gender	-	1.00	0.44	0.33	-0.42
Age	-	-	1.00	0.72	-0.28
Eyeglasses	-	-	-	1.00	-0.19
Smile	-	-	-	-	1.00

(b) Correlation matrix of attribute boundaries from the StyleGAN model trained on FF-HQ dataset.

Tab.4.3 and Tab.4.4 report the results from InterFaceGAN and our proposed method MVU-GAN, respectively. We can clearly observe that our approach is able to identify semantics

Table 4.4: Correlation matrices of attribute directions discovered using our unsupervised method MVU-GAN.

	Pose	Gender	Age	Eyeglasses	Smile
Pose	1.00	-0.07	-0.16	-0.10	0.10
Gender	-	1.00	0.00	-0.02	-0.05
Age	-	-	1.00	-0.12	-0.19
Eyeglasses	-	-	-	1.00	-0.20
Smile	-	-	-	-	1.00

(a) Correlation matrix of attribute boundaries from the StyleGAN model trained on CelebA-HQ dataset.

	Pose	Gender	Age	Eyeglasses	Smile
Pose	1.00	0.09	-0.05	-0.03	0.09
Gender	-	1.00	-0.02	-0.01	0.04
Age	-	-	1.00	0.09	-0.10
Eyeglasses	-	-	-	1.00	-0.12
Smile	-	-	-	-	1.00

(b) Correlation matrix of attribute boundaries from the StyleGAN model trained on FF-HQ dataset.

that present small degree of correlation. Specifically, the gender boundary is almost orthogonal to those of age, eyeglasses and smile for both CelebA-HQ and FF-HQ dataset. In contrast, InterFaceGAN cannot discover well separated boundaries, resulting in high correlation between gender and the other, aforementioned attributes. Also, age and eyeglasses seem to be highly correlated with each other. This is to be expected as it has been already reported from [20] that older men are more prone to wearing glasses. However, our approach presents much less entanglement between these two features. InterFaceGAN is superior to MVU-GAN only in the pose attribute, as our method shows a slightly larger correlation between pose and other feature boundaries compared to InterFaceGAN.

4.2.4 Diversity Study

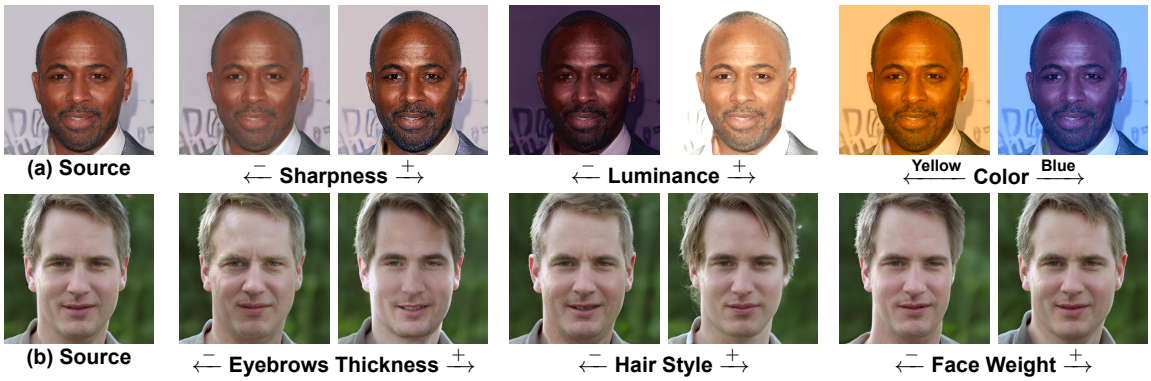

Figure 4.5: (a) Diverse semantics related to color schemes and microstructures, that can *not* be identified by InterFaceGAN [27] due to the lack of semantic predictors. StyleGAN model trained on CelebA-HQ is used. (b) Diverse semantics related to smaller scale facial features, that can *not* be identified by InterFaceGAN [27] due to the lack of semantic predictors. StyleGAN model trained on FF-HQ is used.

As we have already mentioned, the main drawback of supervised approaches is that they rely heavily on the available attribute predictors. Thus, the number of interpretable directions that can be learned is significantly reduced. Fig.4.5 (a) justifies this statement by showing how are approach is capable of discovering semantics corresponding to color schemes and microstructures. The difficulty of obtaining predictors for such attributes forces InterFaceGAN to be deprived of the possibility of discovering such semantic directions. Similarly, InterFaceGAN has been designed to explore more distinctive semantics of larger variation. In comparison, as shown in Fig 4.5 (b), we successfully identify directions corresponding to features of smaller variation, such as eyebrows thickness, hairstyle and face weight.

4.3 Comparison with Unsupervised Approach

As mentioned earlier, our work is primarily motivated by SeFa [28]. In their work, an extensive comparison was made between sampling-based and learning-based baselines and it has been shown that it surpasses both of them. Thus, we choose to compare our framework MVU-GAN with the state-of-the-art unsupervised, independent of any kind of data sampling or model training, method of SeFa. Specifically, we will follow an experimental approach similar to that of Sec.4.2 and will conduct experiments on face synthesis models, evaluating them on the same metrics mentioned above.

4.3.1 Qualitative Comparison

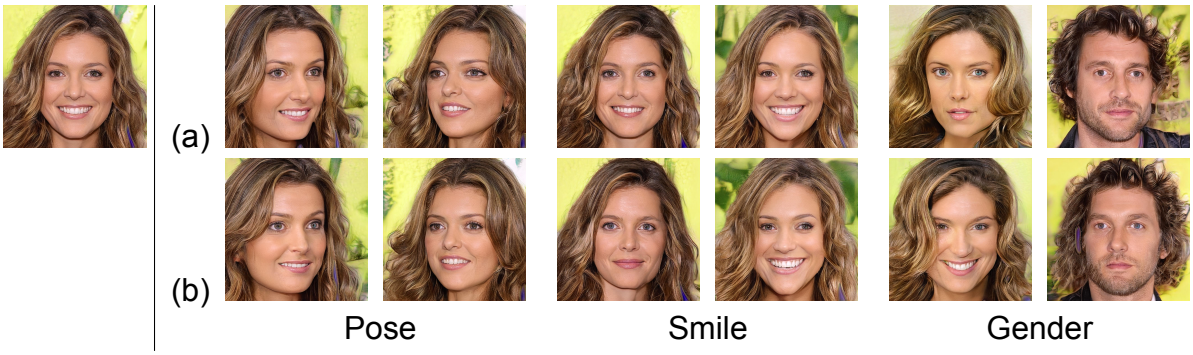

Figure 4.6: Qualitative comparison of the latent semantics found by (a) the unsupervised method, SeFa [28] and (b) our proposed framework MVU-GAN from the StyleGAN model [20] trained on CelebA-HQ [23] dataset.

Fig. 4.6 shows the qualitative comparisons results between SeFa (a) and our method MVU-GAN (b). At first glance, MVU-GAN seems to produce less noisy, more precisely controlled images over a given discovered direction. For instance, the manipulation of gender attribute by SeFa leads to a somewhat distorted image between the person's hair and the background. Moreover, moving towards the negative side of smile direction, SeFa method cannot synthesize a woman who does not smile, as we would normally expect. On the other side, our method MVU-GAN does not suffer from such issues.

4.3.2 Quantitative Comparison

We quantitatively compare our approach with SeFa with FID [12] and re-scoring analysis.

Table 4.5: Quantitative comparison using re-scoring analysis between the unsupervised approach SeFa [28] and our proposed method MVU-GAN for the StyleGAN model [20] trained on CelebA-HQ dataset. Each row quantifies the change of the semantic score after moving into a certain direction in the latent space.

	Pose	Gender	Age	Eyeglasses	Smile
Pose	0.51	-0.14	-0.11	0.01	0.03
Gender	0.02	0.57	0.49	0.08	-0.10
Age	-0.05	0.22	0.39	0.22	0.09
Eyeglasses	0.01	0.55	0.45	0.11	-0.09
Smile	-0.02	-0.03	0.12	0.19	0.30

(a) SeFa

	Pose	Gender	Age	Eyeglasses	Smile
Pose	0.55	0.06	0.12	0.09	0.16
Gender	-0.01	0.65	0.02	-0.03	-0.01
Age	-0.02	0.03	0.45	0.15	0.01
Eyeglasses	0.05	0.19	0.09	0.49	0.03
Smile	0.03	0.01	-0.04	0.00	0.75

(b) MVU-GAN

Table 4.6: Quantitative comparison using FID score between the unsupervised approach SeFa [28] and our proposed method MVU-GAN for the StyleGAN model [20] trained on CelebA-HQ dataset.

Method	FID
SeFa	7.13
MVU-GAN	**6.10**

Tab.4.5 and Tab.4.6 show that MVU-GAN clearly outperforms SeFa both in terms of controllability over the explored directions in latent space and image quality.

4.3.3 Robustness Study

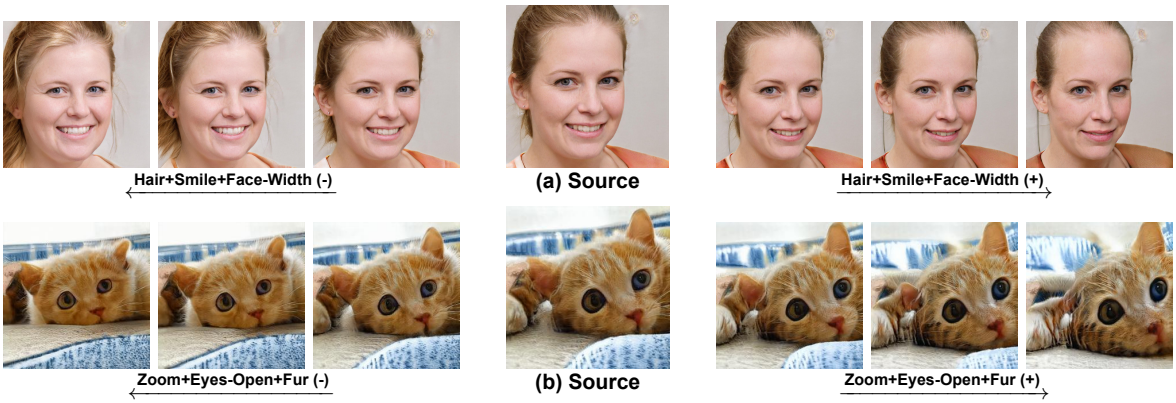

Figure 4.7: Decomposing multiple successive layers from (a) StyleGANv2 model trained on FF-HQ dataset and (b) from StyleGAN model trained on LSUN-Cat dataset, leading to hierarchical multi-attribute manipulation.

Both SeFa and our proposed framework MVU-GAN are unsupervised methods, independent of data sampling and model training. However, we have noticed that when we try to decompose multiple successive layers, SeFa fails to interpret the corresponding pretrained weights, leading to a completely distorted image. Instead, our method is able to decompose them by discovering interpretable directions corresponding to multiple attributes. Specifically, it has been observed that these directions are in practice a combination of the corresponding interpretable directions of the component layers. Fig.4.7 shows our findings for the StyleGANv2 model trained on FF-HQ dataset and for the StyleGAN model trained on LSUN-Cat dataset. In particular, for (a) we decompose the layers $2-7$ while for (b) we decompose the layers $0-10$. We can tell that for FF-HQ dataset only main structural changes occur (hair, smile, face width). That is because we omit bottom layers $(0-2)$ from this specific decomposition, which correspond to high level aspects of the image, such as pose etc. On the other hand, for LSUN-Cat dataset where bottom layers are decomposed along with the middle and the top ones, we observe changes in every hierarchical set of layers, e.g. zoom corresponds to the bottom layers, the opening of the eyes in the middle layers and the fur change in the top ones.

5. CONCLUSIONS

In this thesis, we propose an innovative unsupervised method, MVU-GAN, to interpret the latent space of pretrained GANs. Extensive experiments demonstrate that our approach outperforms the existing state-of-the-art ones. Specifically, we show that our algorithm is capable of discovering a wide variety of interpretable directions, as well as different variations of a certain attribute. The proposed method has also been shown to be robust in decomposing multiple successive layers, leading to high-quality multi-attribute manipulation.

ABBREVIATIONS - ACRONYMS

GAN	Generative Adversarial Network
MVU	Maximum Variance Unfolding
SDP	Semidefinite Programming
CNN	Convolutional Neural Network
MLP	Multilayer Perceptron
PCA	Principal Components Analysis
KNN	K-Nearest Neighbor
Flickr-Faces-HQ	FFHQ
Large-scale CelebFaces Attributes	CelebA-HQ
Perceptual Path Length	PPL
Fréchet Inception Distance	FID
Inception Score	IS

BIBLIOGRAPHY

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017.

[2] Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: on the curvature of deep generative models, 2018.

[3] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks, 2018.

[4] David Berthelot, Thomas Schumm, and Luke Metz. Began: Boundary equilibrium generative adversarial networks, 2017.

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis, 2019.

[6] Nutan Chen, Alexej Klushyn, Richard Kurle, Xueyan Jiang, Justin Bayer, and Patrick van der Smagt. Metrics for deep generative models, 2018.

[7] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets, 2016.

[8] Ishan Durugkar, Ian Gemp, and Sridhar Mahadevan. Generative multi-adversarial networks, 2017.

[9] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties, 2019.

[10] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[12] 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, 2018.

[13] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization, 2017.

[14] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls, 2020.

[15] Ali Jahanian, Lucy Chai, and Phillip Isola. On the ”steerability” of generative adversarial networks. In *International Conference on Learning Representations*, 2020.

[16] Yanghua Jin, Jiakai Zhang, Minjun Li, Yingtao Tian, Huachun Zhu, and Zhihao Fang. Towards the automatic anime characters creation with generative adversarial networks, 2017.

[17] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan, 2018.

[18] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation, 2018.

[19] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks, 2021.

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019.

[21] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan, 2020.

[22] Line Kuhnel, Tom Fletcher, Sarang Joshi, and Stefan Sommer. Latent space non-linear statistics, 2018.

[23] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[24] Antoine Plumerault, Hervé Le Borgne, and Céline Hudelot. Controlling generative models with continuous factors of variations, 2020.

[25] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016.

[26] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans, 2016.

[27] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans, 2020.

[28] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans, 2021.

[29] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space, 2020.

[30] Jianzhong Wang. *Maximum Variance Unfolding*, pages 181–202. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[31] Ceyuan Yang, Yujun Shen, and Bolei Zhou. Semantic hierarchy emerges in deep generative representations for scene synthesis, 2020.

[32] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop, 2016.

[33] Yang Yu, Zhiqiang Gong, Ping Zhong, and Jiaxin Shan. Unsupervised representation learning with deep convolutional neural network for remote sensing images. In Yao Zhao, Xiangwei Kong, and David Taubman, editors, *Image and Graphics*, pages 97–108, Cham, 2017. Springer International Publishing.

[34] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks, 2019.