DOI: 10.1111/cgf.14524
Eurographics Conference on Visualization (EuroVis) 2022
R. Borgo, G. E. Marai, and T. Schreck
(Guest Editors)

Volume 41 (2022), Number 3

Interactively Assessing Disentanglement in GANs

Sangwon Jeong[1], Shusen Liu [2], and Matthew Berger[1]

[1]Vanderbilt University, USA
[2]Lawrence Livermore National Laboratory, USA

Abstract

Generative adversarial networks (GAN) have witnessed tremendous growth in recent years, demonstrating wide applicability in many domains. However, GANs remain notoriously difficult for people to interpret, particularly for modern GANs capable of generating photo-realistic imagery. In this work we contribute a visual analytics approach for GAN interpretability, where we focus on the analysis and visualization of GAN disentanglement. Disentanglement is concerned with the ability to control content produced by a GAN along a small number of distinct, yet semantic, factors of variation. The goal of our approach is to shed insight on GAN disentanglement, above and beyond coarse summaries, instead permitting a deeper analysis of the data distribution modeled by a GAN. Our visualization allows one to assess a single factor of variation in terms of groupings and trends in the data distribution, where our analysis seeks to relate the learned representation space of GANs with attribute-based semantic scoring of images produced by GANs. Through use-cases, we show that our visualization is effective in assessing disentanglement, allowing one to quickly recognize a factor of variation and its overall quality. In addition, we show how our approach can highlight potential dataset biases learned by GANs.

CCS Concepts
• ***Computing methodologies*** → ***Artificial intelligence; Model verification and validation;***

1. Introduction

Interpretability is increasingly becoming an important problem for the adoption of machine learning models in a variety of applications. For discriminative models, the ability for humans to interpret the rationale behind a model decision, e.g., why a prediction was made in a classification task, can help instill trust in users when applying the model downstream [CL18, DBH18]. The manner in which humans consume interpretations of discriminative models is typically passive: the model makes a prediction, and then we provide some explanation for its prediction. In contrast, for generative models [GPAM*14, Jeb12], the problem of interpretability is more complex, as end-users tend to be more active in using these models, e.g., generative models are maturing as a mechanism for design, ideation, and content creation within a number of domains [PvdWR6, KALL17, ZXL*17, LTH*17, MSI*18, FADK*18, TCAT17, Mog16]. For these use cases, the ability to understand the latent semantics within a generative model is of high importance as it can benefit how one ultimately controls the model for the target applications.

Identifying distinctive semantics, i.e., disentanglement, in the generator is the predominant approach to making sense of a generative model. Disentanglement is concerned with the discovery of directions, found within the latent space of a generative model, where a single direction tends to capture a semantic, human-nameable concept, and only that concept. Such disentangled directions are ideal for controlling generative models, especially considering that a generator's latent space is usually a dense, high-dimensional vector space that is otherwise challenging to interpret. Many semantic discovery approaches have been proposed, e.g., ensuring individual components of the vector space capture distinct concepts [HMP*16], supervised direction-finding given concept-annotated datasets [SYTZ20, YSZ21], as well as unsupervised direction-finding methods [VB20, SEBM20, HHLP20, SZ20].

Despite the rapid progress, a fundamental issue with existing works is in evaluation: it is challenging to quantify just how disentangled a given direction is, particularly for modern models [BDS18, GAOI19, KALL17, KLA19]. Conventionally, summary statistics are proposed based on a predefined characterization of disentanglement, e.g., Eastwood & Williams [EW18] propose completeness, informativeness, and distinctiveness as desirable properties, alongside heuristic measures for each. On the other extreme, individual draws from the latent space can be shown to users for validation [HMP*16]. Though we see value in both, we believe that a middle ground exists for improving our understanding of latent space directions, where interactive visualization can support the user in evaluating a given direction.

To this end, we propose a visual analytics approach for assessing disentanglement within generative models. Our visualization

Figure 1: *Problems of entanglement within GANS: walks in the latent space are generated from three different latent codes, but along a "youth" direction. Note the different changes along this direction: the first walk does not change significantly in any other attributes but age, but the second walk does change in smiling. Eyeglasses disappear along with age in the last example.*

design is centered on **walks** in the latent space of a generative model. Specifically, a walk is defined by taking a single **code** (a high-dimensional vector) in the latent space, and producing a sequence of samples that lie along a given **direction** (another vector). At a high level, our goal is to help users understand the quality of a given direction, through visually analyzing a collection of walks, the result of taking a collection of latent space codes and producing a walk for each code. Our approach relies on a set of pre-trained image classifiers that produce continuously valued scores, indicating the extent to which an image contains a particular attribute or concept that can be easily recognized. For a single walk, and its corresponding generated images, we can obtain a sequence of scores for each attribute.

This detailed attribute-based information forms the basis for our visualization design and various analyses that we aim to support. Specifically, if there exists a consistent trend across walks for a given attribute, e.g., the attribute scores monotonically increase/decrease, we believe the direction is representative of the attribute. On the other hand, there may exist attribute correlation with respect to the trends, e.g., for face images, attributes representative of smiling and cheekbone are likely to be correlated - when one increases, the other is likely to increase as well. Identifying when correlation exists, and whether or not the correlation is sensible, is critical for determining the quality of a direction, above and beyond merely identifying distinct factors of variation. Furthermore, not all codes in the latent space need to have an equivalent response to the direction, giving rise to different attribute trends. This particular matter is exemplified in Figure 1.

Determining groups of walks that have similar trends thus becomes essential to obtain a more comprehensive understanding of the direction. Our visualization design is intended to help the user probe a given direction in support of these types of analyses, through a set of linked and coordinated views that depict attribute scores, their trends, and more general similarity of walks.

We conduct use cases through visually exploring StyleGAN [KLA19] using a generative model of high-resolution faces, alongside facial attribute classifiers [SYTZ20]. Our interface allows for the analysis of arbitrary directions, where we investigate directions produced from supervised methods [SYTZ20], unsupervised methods [SZ20], as well as simple, vector arithmetic of latent codes. Although we present our work only on facial attributes, this approach is generalizable to any image domain with a GAN model and attribute classifiers. For instance, a GAN optimized to generate images of natural scenery or places [ZLK*17] can be evaluated using attribute classifiers trained using a transient attribute dataset [LRT*14]. We contribute to two main use cases, enabling a deeper understanding of directions in generative models:

- We show directions computed from unsupervised and supervised methods that capture, at face value, the same predominant semantic property, nevertheless differ significantly in other properties.
- We demonstrate an exploration of bias in directions, wherein continuous semantics that are correlated across walks highlight a learned bias in the model.

2. Related Work

There are several areas of research that are relevant to the proposed approach. Major bodies of related work revolve around semantic discovery in latent space that mostly originates from the machine learning community, as well as a variety of visual analytic approaches for exploring latent spaces that arise from the visualization community.

2.1. Supervised Semantic Discovery

The ability to identify and explore high-level concepts in latent representations is crucial for making sense of neural networks. Often, we can identify a vector direction in the latent space that corresponds to clear semantics. From the early discovery that semantics can be obtained through vector arithmetic in word embedding spaces [MSC*13], to later work that confirmed meaningful concepts can also be encoded as a linear direction in the latent space of GANs [RMC15] and CNNs [KWG*18], a variety of approaches have been proposed to uncover these vector direction through supervised and unsupervised means. Supervised methods typically rely on external models or known transformations to identify meaningful directions. In TCAV [KWG*18], example images of the concept of interests are gathered in order to train a simple linear model to identify the concept direction in the latent space. The "interface" work [SZ20] employs an external model that assesses attribute scores of an image. The semantic directions are derived from SVM classifiers that were trained on the GAN-synthesized random images using labels generated by a pre-trained classifier. Ganalyze [GAOI19] employs an external model that assesses the memorability of the generated image. They find directions such that for a given starting image, a manipulated image's memorability score results in a large change. However, they do not consider a wider variety of different semantics. The GAN steerability work [JCI19] finds a direction by first sampling an image using a randomly sampled latent code, applying an edit such as zoom or rotation to the generated image, and recording a latent code that produces an image that minimizes the loss between itself and the previous image. However, the method can be computationally heavy and only considers coarse image manipulations.

2.2. Unsupervised Semantic Discovery

Despite the effectiveness of supervised methods, they are limited by the type and variety of concepts that can be identified, and more often than not we may not have labels to utilized the supervised method to begin with. Voynov et al. introduce an unsupervised approach [VB20], relying on the intuition that meaningful concepts are often more disentangled with each other and therefore more predictable. The method trains a special component called "reconstructor" that takes two images generated by latent codes along a vector direction (a column in a matrix that is part of the trainable parameter). and predicts which vector/concept induces the difference between the images and the magnitude of the change. GANSpace [HHLP20] identifies meaningful directions in a GAN's latent space by applying PCA to latent codes in subsequent layers, subsequently mapping them back into the original latent space. Shen et al. [SZ20] propose a closed-form approach, where decomposition of a given model's weight matrix is used for identifying directions that result in large changes in the model, and consequently, in the output images. Among both these supervised and unsupervised approaches, many introduce metrics for evaluating the quality of obtaining directions that are based on ground truth obtained from supervised methods, the optimization objectives, and human evaluation. However, all these approaches focus on producing a single number instead of providing more granularity in their evaluation, i.e., where and how did some of these methods fail, which can be crucial to uncover the cause of the failure and possible avenues for improvement. The proposed work aims to address this gap in existing works.

2.3. Disentanglement in Latent Space

Most semantic discovery methods approach the problem from a post hoc perspective, however, we can also directly build a latent space in which each dimension has a distinctive concept. Many approaches have been proposed, such as beta-VAE [HMP*16] and InfoGAN [CDH*16], for obtaining a disentangled latent space. However, defining and estimating whether different factors are disentangled is a non-trivial task. In [HAP*18], Higgins et al. aim to work toward a formal definition. Moreover, the inherent bias in the data may also contribute to challenges in understanding disentanglement, and the fairness of disentangled representations is questioned [LAR*19]. One analysis goal of the proposed method is to provide a visual analytic centric alternative for examining and understanding the disentanglement between the discovered semantic directions, whether it is from a supervised or unsupervised method (or even a disentangled latent dimension).

2.4. Visual Exploration of Latent Space of GAN

Besides methods focusing on purely computational methods, visualization approaches provide powerful and flexible alternatives for exploring concepts in latent spaces. In the latent space cartography [LJLH19] work, Liu et al. introduced a visualization system that maps and compares meaningful semantic dimensions within latent spaces. For natural language processing, the word embedding latent space are explored by several visualization works that focus on various aspects, from examining semantics encoded in vector directions [LJLH19] to comparing the embeddings themselves [HKMG20]. Several visual analytics approaches specifically focused on visualizing GAN and its training process. The GANlab [KTC*18] provides a playground-like environment for understanding how GANs work. By leveraging a 2D function estimation problem, the system allows user to see how GANs evolve during training and explore how different hyperparameters affect the output. GANViz [AKBR19] aims to shed light on the complex training dynamic among different components of a GAN, to help domain experts evaluate and potentially improve their models. Our work contributes towards similar, higher-level aims of understanding GANs, but is instead focused on assessing directions in GAN latent spaces.

3. Objectives & Tasks

The central goal of our work is to support users in obtaining a deeper understanding of disentanglement, provided a direction within the latent space of a generative model. We want to help users assess how good the direction is, as an avenue for generating meaningful factors of variation, provided an arbitrary code in the latent space. Forming an accepted, understood definition of disentanglement remains challenging; nevertheless, a number of works have introduced properties of disentangled representations [EW18, HAP*18, LAR*19], and corresponding heuristics for measuring these properties. We are inspired by these prior works, but instead approach evaluating disentanglement from an exploratory perspective.

More specifically, we would like the user to obtain a better understanding of disentanglement through assessing the following properties:

- **[O1] Distinctiveness** : how many factors of variation exist for a given direction. A greater variation tends to indicate the direction is not distinct.
- **[O2] Consistency** : how predictable the variation is for *any* latent code. A direction is considered consistent if the same factors of variation are present across codes in the latent space.
- **[O3] Informativeness** : the extent to which an end-user can recognize, e.g. easily name, what the direction represents.

Limiting our analysis to only the generator, the representations that it learns, and its image output, poses significant challenges for addressing the above objectives. This is due to the complexity of the data, once provided a direction, e.g. for a walk we have a sequence of latent codes and their corresponding generator representations/images. To support our analysis, we rely on **attribute** classifiers, ones that score an image for a prescribed, semantic concept, e.g. various facial attributes. The use of attributes enables us to simplify a walk in the latent space to a sequence of attribute scores, similar to prior work [SZ20]. However, in our work we aim to use attribute classifiers, alongside the generator representations, to obtain a more comprehensive understanding of disentanglement. Given this information, our visualization design is driven by the following set of user tasks aimed at satisfying the above objectives:

- **[T1]** Distinguish salient attributes, e.g. those that have similar trends across codes, from less salient attributes. A small number of salient attributes suggests a distinct direction. **[O1]**

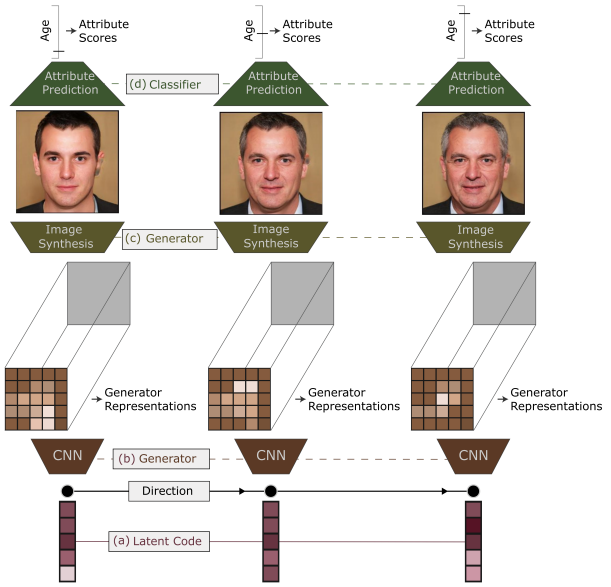

Figure 2: *Overview of the data generation process for a walk of a single latent code, taking three samples along the walk: (a) a sequence of latent codes are input to the generator, (b) from which we extract convolutional activations as a generator representation of the code. Next, generated images (c) are assessed by external attribute classifiers (d), subsequently used to explain the walk in terms of classified attributes.*

- **[T2]** Identify trends in attributes with respect to walks, e.g. for a given walk, which attributes monotonically increase/decrease in score. Such trends signal consistency in directions. **[O2]**
- **[T3]** Analyze correlations between attributes. A set of attributes might change in consistent ways, but may be unrelated to one another, indicative of uninformative directions. **[O3]**

4. GAN Latent Space Analysis

In this section we discuss the generator model that we study in our work, as well as corresponding data we extract from the generator for visual analysis.

We focus our analysis on the StyleGAN [KLA19] model, due to its ubiquity within the machine learning community, and its capability to produce realistic, high-resolution images. StyleGAN synthesizes images by first drawing a high-dimensional vector, whose components are independently sampled from a (truncated) normal distribution, and feeding this vector through a multi-layer perceptron (MLP), yielding what is colloquially known as the $\mathcal{W}$ space. Unless otherwise stated, this latent space is the focus of our analysis, where we denote a sample in the latent space as a **code**, a d-dimensional vector (c.f. Figure 2(a)). The code is fed through a series of convolutional layers (c.f. Figure 2(b)), whose activations are modulated by style-based features, to ultimately produce an image (c.f. Figure 2(c)).

We are further provided a direction in the latent space, denoted as a unit-norm vector $\mathbf{a} \in \mathcal{W}$. Given a code $\mathbf{z} \in \mathcal{W}$, we sample a sequence of m equal-spaced steps along the direction, centered at a direction-neutral transformation of the code, denoted $\tilde{\mathbf{z}} \in \mathcal{W}$, found by projecting the original code onto the orthogonal complement of the direction, namely, $\tilde{\mathbf{z}} = (I - \mathbf{a}\mathbf{a}^\top)\mathbf{z}$.

This provides us a sequence of codes $\mathbf{z}(t) \in \mathcal{W}$ where t indexes over steps. We denote this sequence as a walk in the latent space, and we may obtain a sequence of images by feeding each code through the rest of the StyleGAN model. Ideally, the further that a code lies along a direction $\mathbf{a}$, relative to the neutral code $\tilde{\mathbf{z}} \in \mathcal{W}$, then a single, recognizable concept will become more prominent in the image. The converse should also hold for codes that point in the opposite direction, e.g. if a direction captures the facial property "closed eyes", then such codes result in faces with eyes wide open.

We are interested in understanding a collection of walks, formed by sampling codes in the latent space, and for each, generating a sequence of samples as described above. To aid in the analysis of walks, we assume the existence of a set of pretrained classifiers optimized to predict the presence of attributes - semantic, recognizable concepts that describe an image. We use the raw outputs of these pretrained models, rather than their binarized classification decisions, to obtain continuously-valued scores and thus a means of quantifying the presence/absence of an attribute. Thus, for each code along a given walk, we take its synthesized image and run the attribute classifier to obtain a score for this image (c.f. Figure 2(d)).

Although informative, a sequence of raw scores for each attribute is not the most concise way of describing a direction's relationship with respect to a given attribute. This is more so when analyzing a large collection of walks and comparing them across attributes. Thus, we summarize a sequence of attribute scores by fitting a linear regression model to a given walk's sequence of attribute scores. Our assumption is that attribute scores will largely either monotonically increase, decrease, or remain constant, as the walk progresses along a direction, indicating a salient change (or lack of change) in the attribute. Specifically, the slope of the resulting model suggests the type of trend, while the R^2 coefficient indicates the goodness of fit (confidence) - we use both of these values to summarize a sequence of attribute scores.

Last, in order to ensure that our analysis is not biased by the predefined collection of attributes, we would also like to have a more general notion of similarity between walks, in order to better understand their distribution. To this end, we extract representations from the generator as a proxy for similarity - indeed, most (un)supervised walk finding methods are based on features of the generator, and so we expect this to encode semantically-meaningful information. Specifically, given a walk, for each of its codes, we take the convolutional output at a given layer in the generator, typically a layer somewhat close to the latent space , in order to complement the attribute classifiers which operate on images. We further justify our decision on generator layers in supplemental material. For each output we perform max spatial pooling to obtain a series of vectors $[v_0, v_1, ..., v_{t-1}]$ where t is a number of latent codes in a walk. Then vectors along the walks are subtracted $[v_1 - v_0, v_2 - v_1, ..., v_{t-1} - v_{t-2}]$. Subtraction is followed by concatenating all of the difference vectors. Concatenation yields a feature representation for the walk. Last, we perform UMAP

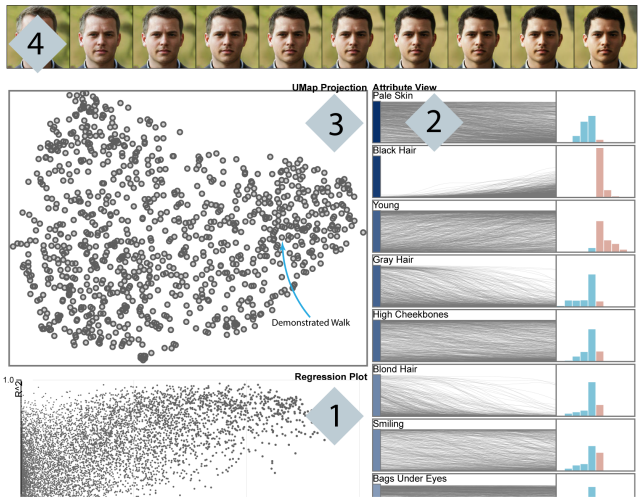

Figure 3: *An overview of our visualization interface. (1) The regression plot shows the absolute slope and R^2 coefficient of the sampled walks' attribute scores. At-a-glance, this helps convey whether a direction is worth investigating in detail. (2) The attribute view shows a summary of selected walks on each attribute. Users can assess relationships between attributes via this view. (3) The UMAP projection shows the structure of sampled latent code walks. This is useful in identifying and selecting clusters of walk that correlate to an attribute differently e.g., has a negative correlation when most of the samples have a positive correlation. (4) On-demand generated images are available for verification i.e., whether the assessment made by attribute classifier models are indeed correct. This can instill user trust in the visualization.*

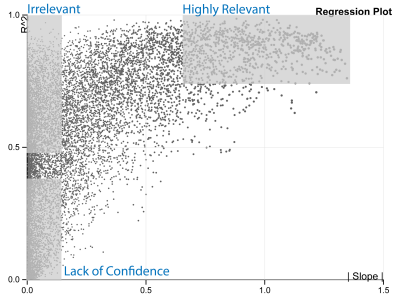

Figure 4: *Regression scatterplot: here users can assess if a direction is worth investigating i.e, there exist numerous highly-relevant and confident walks. Brushing is supported to select walks, linked and coordinated with the attribute view to support further investigation.*

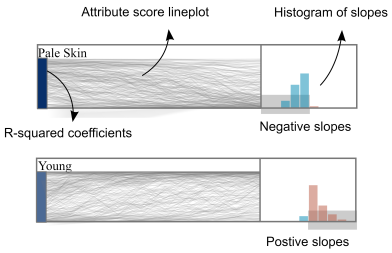

Figure 5: *We allow the user to see, for a given attribute, a summary of slopes over linear regression fits as a histogram, alongside detailed attribute scores for all walks.*

[MHM18] to obtain a 2D projection for the set of walks, e.g. each 2D position will correspond to a single walk.

5. Visualization Design

Our visualization is designed for users to interactively explore a provided direction, and assess disentanglement, built around the tasks and data described above. Our design is comprised of four main views (c.f. Figure 3): (1) a detailed view of linear regression models, (2) a depiction of attribute scores and summary of attribute trends, (3) a general similarity view on walks, and (4) detailed inspection of individual walks.

5.1. Visual Encodings

Regression Plot [Figure 3-1]: Within the scatterplot, the points correspond to all combinations of walks and their attribute-specific linear regression models, where we encode the model's slope on the x-axis and its R^2 coefficient on the y-axis. This allows the user to see, at-a-glance, the quality of a selected direction (c.f. Figure 4). For instance, if all points are concentrated around the centerline, then this suggests the direction is not reflective of any of the provided attributes. On the other hand, points that spread out near the upper right of the plot suggest a subset of walks that consistently decrease or increase along with certain attributes.

Attribute View [Figure 3-2]: We draw score of walks for each attribute as a single line in the attribute view. Next to the line plots, we show a summary of the slopes as a histogram. Here, positive slopes are colored as red bars and negative slopes as blue. Attributes are sorted in a descending order of R^2, so most relevant attributes for a given direction can be quickly determined. Please see Figure 5 for an illustration.

UMAP Projection [Figure 3-3]: We show UMAP projection of walks' generator representation in this view. The view's role is to act as a window for selecting walks in terms of similarity in their activations. This representation of the walks is decoupled from the information that is generated through attribute classifiers, which allows for the cross-examination of walks using two different descriptions of walks.

Generated Images [Figure 3-4]: Last, we let users to inspect walks in detail by showing a walk's corresponding sequence of synthesized images. Inspecting actual images is a confirmation of findings made in other views - generator representation and attribute-based quantitative information. Showing a sequence of images can reveal whether a selected walk indeed captures expected changes with respect to the identified attributes. Reassurance gives user confidence, which is a by-product of revealing detailed information.

When users choose to visualize a large number of samples, the visualization can suffer from over-plotting. Although over-plotting

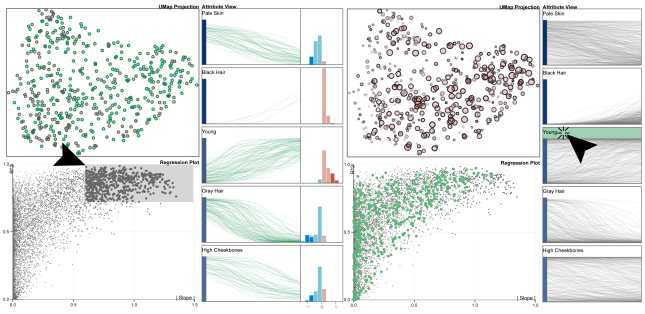

Figure 6: *The figure shows three main interactions that a user can perform. A user can brush on a UMAP projection and a regression plot to view the selected walks' attribute scores and their corresponding slope in a form of a histogram. Hovering over attribute labels highlights the corresponding attribute regressions in the scatterplot and reveals each latent code's correlation with the attribute in the Umap projection. Clicking on an attribute label locks other plots with respect to the clicked attribute, this prevents users from having to memorize other views.*

can hinder readability of some views, we chose our current encodings for several reasons. First, line marks in the attribute view could be substituted by an averaged trend. However, this will lead to a loss of information such as the detection of outliers. Second, a regression plot could be represented as a density plot. However, dense regions would need to be highlighted using an appropriate color channel, introducing interference with our remaining views and interactions, e.g. using brightness would interfere with our choice of brushing. Therefore, we chose to preserve the individuality of points to let them stand out amidst our supported interactions, which we detail next.

5.2. Interactions

The individual views are linked and coordinated through a set of interactions that users may perform (c.f. Figure 6). Our visualization is broadly designed to be exploratory, and effective workflows can differ from person to person depending on their goals. Here we detail our interaction design, and illustrate how the interactions may be used in practice.

Upon selecting a direction to investigate, users can brush over the Regression Scatterplot. Brushing results in the selection of a subset of walks $S \subset D$ that are within a range of slopes and goodness-of-fits of interest. Note the set S is comprised of pairings of latent codes with their attributes, those that satisfy the brushed query. In response to the brush, we update the Attribute View, limiting the line marks to those in S, and superimposing the histograms with darker-hued bars reflecting the counts arising from the selection S. We further update the UMAP Projection, highlighting a point if it belongs to a code in S, regardless of its subset of attributes. We view this brushing action as a suitable starting point for analysis, e.g. one can select codes and attributes that have high R^2 coefficients to observe trends in the attribute score line plots, thus allowing users to assess the distinctiveness and informativeness of the selected direction **[T1, T3]**

Next, users can either hover or click on the attribute text labels in the Attribute View. Both actions have the same effect where clicking locks attribute selection as opposed to a non-persistent effect from hovering. Linked updates are made to both the Regression Scatterplot and UMAP Projection. In the Regression Scatterplot, all walks that correspond to the selected attribute will be highlighted – this helps to contextualize the selection (S) focus. In the UMAP Projection, walks not present in the selected attribute in the regression plot selection will be filtered out. Those walks whose latent codes are present in S, and limited to the selected attribute, will be updated by using the slope and R^2 coefficients. Here, a magnitude of a slope is encoded by the radius of circles and the slope's sign encoded by the color, consistent with the color scheme of the histogram. R^2 coefficients are encoded via stroke thickness of circles. Through selecting attributes, consistency of the direction can be studied - to verify whether groups of walks that are in close proximity in the UMAP Projection also have related, and confident, trends for the chosen attribute **[T2]**.

After choosing an attribute, users can brush on the UMAP Projection, prompting linked updates back to the regression and attribute views in accordance with selected walks across all attributes, de-aggregating the previous selection from the Regression Scatterplot. In turn, the Attribute View is updated for a new subset of walks $S' \subset D$ that resides within the brushed area of the UMAP projection. This interaction is useful for determining whether groups of walks that have a similar generator representation are consistent in their attribute score **[T2]**.

Last, clicking on a point in the UMAP projection will bring up the sequence of synthesized images that correspond to the selected walk. Based on findings made through using the interface, inspecting individual images of walks can support the user in verifying hypotheses they might have formed through the prior sequence of interactions.

6. Experimental Results

We show how our visualization can be used to perform a variety of analyses for better understanding directions in GANs. Our current interface supports a StyleGAN model optimized on the FFHQ dataset [KLA19], and use the CelebA dataset [LLWT15] for attribute analysis. However, we note that the interface can be used on any modern GAN model with attribute classifiers. Our interface is available via the link `https:observablehq.com/d/58f90c0a153bd534`.

Our interface supports three different types of direction finding methods:

- **Vector arithmetic**: we manually construct directions by identifying a pair of latent codes whose images differ by a semantic concept, and take their difference. For robustness, we average three latent codes with the attribute and another three without the attribute.
- **Unsupervised direction finding**: we use the approach of SeFA [SZ20] to find directions directly from the StyleGAN style layer mappings, e.g. directions that result in large change in the linear maps. This provides us with directions that depend on (a) the

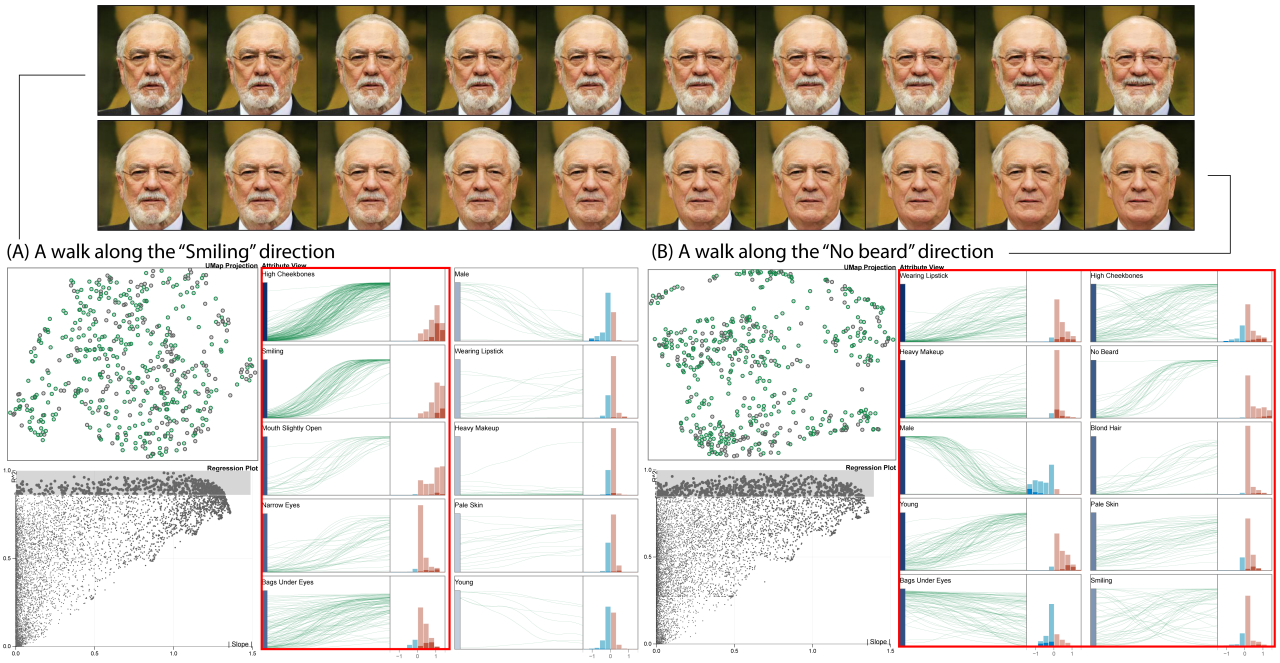

Figure 7: *In this figure we compare two directions to contrast their distinctiveness. (A) is a walk along the "Smiling" direction, and it is associated with the top image row. (B) is a walk along the "No Beard" direction, and it is associated with the bottom image row. Attributes that are highly relevant are the ones with a red bounding box. It can be said that the "Smiling" direction is more distinctive than the "No beard" direction because it is related to less attributes.*

layer under investigation and (b) their overall influence on the style layer.

- **Supervised direction finding**: we use the approach of Shen et al. [SYTZ20] to find directions in the latent space that are discriminative of predefined attributes, measured in the corresponding synthesized images. The attributes we use for analysis are also used for supervised direction finding.

In our analyses we compare directions found for a given supervised method, comparison between different direction-finding methods, and assess potential bias that exists within a given direction

6.1. Direction Assessment

We assess directions with respect to the main properties of disentanglement: distinctiveness, consistency, and informativeness.

We first show how to verify whether a direction is distinct and consistent. We expect a direction to be distinct if its factors of variation can be described by a small set of attributes, and for it to be consistent if most attribute scores monotonically increase/decrease across all codes. To help us understand these properties, we take two directions found through the supervised direction-finding method that are based on "Smile" and "No Beard" attributes.

Figure 7 shows a comparison between these two directions. Within the Regression Plot, we brush all walks that have a high R^2 coefficient, in order to limit our analysis to regression fits that are confident. Through this interaction, we observe that the "Smile" direction can be described by a small number of attributes, whereas the "No Beard" direction results in a larger number of attributes (see annotated red rectangles), indicating that the "Smile" direction is more distinct. Furthermore, consistency of the "Smile" direction can be verified via the linked highlight of the histogram plots, indicating a high concentration of large slope values across the most confident regression fits - this is further indicated by the group of points in the upper-right quadrant of the Regression plot, namely those fits that are confidently increasing.

Figure 9 further verifies the "Smiling" direction for informativeness, or whether the set of attributes related to the direction reflect the dominant factor of variation. Histograms highlighted for "High Cheekbones", "Smiling", "Bags Under Eyes", and "Narrow Eyes" indicate a correlation amongst these attributes, and these attributes indeed tend to change as one smiles. However, we note that the strength of correlation between certain attributes differ. Indeed, the "Bags Under Eyes" and "Narrow Eyes" attribute shows a smaller correlation with the brushed set of examples, indicating that this direction mixes less with attributes that do not correlate with the semantic that it captures.

Outside of verifying directions found through supervised methods, we can also use our interface to assess directions found through simpler schemes, namely vector arithmetic. In Figure 10, we have built a direction based on the extremes of smiling i.e., a bag of latent codes that generates bigger smiles and a bag of codes that generates frowns. Directions built from vector arithmetic are more suscepti-

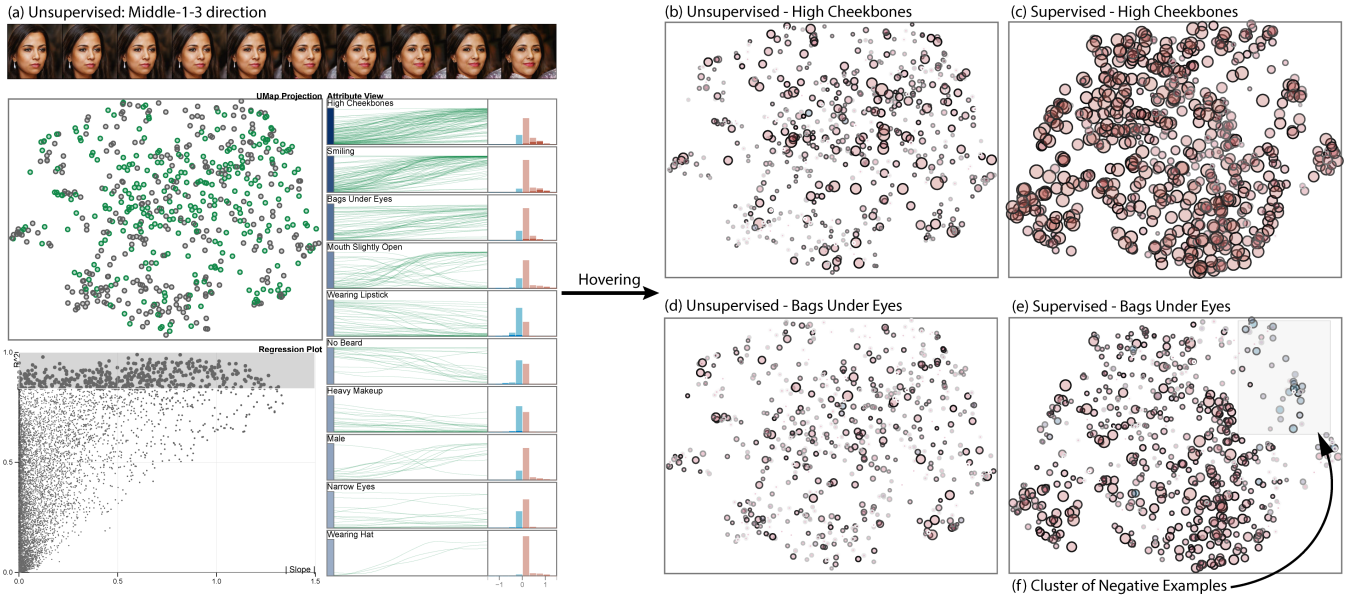

Figure 8: *Comparing direction finding methods: in (a) we show the basic procedure for assessing a direction. Hovering over the cheekbone attribute prompts a response to the scatterplot in (b) and (c), while hovering over the smiling attribute prompts a response in (d) and (e). We can immediately observe that the supervised direction finding method results in far more consistent trends, encoded as the sizing of points int he scatterplot. However, we can also spot some outliers in the supervised method for certain walks that behave differently (blue circles – decreasing trend).*

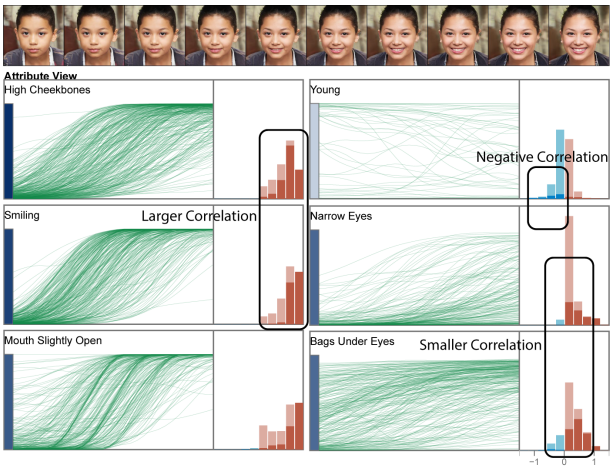

Figure 9: *Informativeness of the "Smile" direction. We find that a "Smile" direction results in correlation of attributes that can be easily understood. However, the direction may not be informative for a small set of codes, e.g. shown here as the "Young" attribute can also consistently decrease for certain portions of the latent space.*

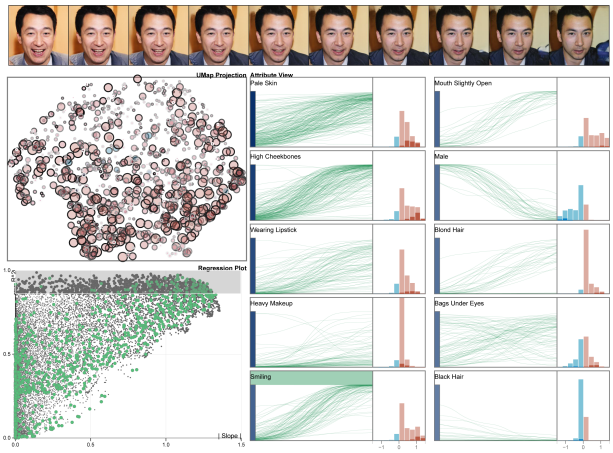

Figure 10: *We show the types of variations produced when using a direction corresponding to extreme smiling, built from vector arithmetic method. We see a very similar correlation to attributes that were seen with the "Smiling" direction built from the supervised method.*

ble to larger variation due to the more localized, hand-crafted approach. Hence, we see that the direction is less distinctive – more attributes' involved with the direction. Simply comparing related attributes in Figure 10 and Figure 7 (A), we can see that the latter is more distinctive.

6.2. Comparing Direction Finding Methods

Here we study the differences that may exist between direction-finding methods, given the same predominant variation found in both. Namely, we compare the "Smiling" direction found through the supervised and unsupervised method. For the unsupervised

method, we manually selected the direction via manual inspection of directions found at different layers.

Figure 8(a) shows our interface for the direction found via the unsupervised method. In "Middle-1-3" direction, middle-1 means that the direction is found in the second convolutional block of StyleGAN generator, the last number 3 means that the direction corresponds to a fourth singular vector. In this direction, we observe consistent trends in "High Cheekbones", "Smiling", and "Bags Under Eyes" for confident regression fits (see: brush in the Regression Plot), further supporting our identification of this direction as "Smiling". Additionally, we find that "High Cheekbones" and "Bags Under Eyes" are attributes that have confident fits in the supervised direction as well. Hence, we use both of these attributes as a basis for comparison (Figure 8 - right). Specifically, we highlight these attributes for both directions, and show the resulting scatterplots.

From here, we can draw two conclusions. First, the supervised method finds a more salient direction for smiling, given the correlated attributes have more consistent trends (Figure 8(c)) in comparison with the unsupervised method (Figure 8(b)), e.g. walks in the UMAP projection tend to have a larger, red circle with bold stroke, indicative of confident and increasing trends. Secondly, although the unsupervised direction is less salient, the grouping in the UMAP projection is helpful for identifying what portions of the latent space give us a salient change in smiling, e.g. the group of points in the center-right portion of Figure 8(b) correspond to children smiling.

The supervised direction is not without its limitations, however. In Figure 8(d) we can observe a small amount of walks which contain a decrease with respect to "Bags Under Eyes". Upon closer inspection, we find these walks introduce glasses as the presence of smiling increases. These types of walks are not present in the unsupervised direction.

6.3. Bias in the Latent Space

Since GANs learn to generate images that follow the data distribution of the provided dataset, they are prone to biases present in the dataset. Our interface can be used to study these biases - how semantics are represented in the latent space. We study bias using two supervised directions - age and pale skin - the interactive example is available through links given in the figure description.

Bias can be studied by examining consistent attributes for a direction. Therefore, we use the same interaction used in assessing a direction - brushing the regression plot in a well-fit region. Once we identify consistent attributes, we can reason whether these attributes can be explained.

Figure 11 shows bias for the "Young" direction. We can observe several attributes that are unrelated to youth, e.g. "Wearing Lipstick", "Male". Indeed, we find that as we walk along the young direction, many walks end up with female gender. This suggests a gender bias in the dataset: more older men than older women. Further, we also find that older people end up with glasses along this direction. A direction that exclusively varies along age should not be entangled with glasses, further suggesting a bias in the dataset wherein older people tend to wear eyeglasses for corrective vision.

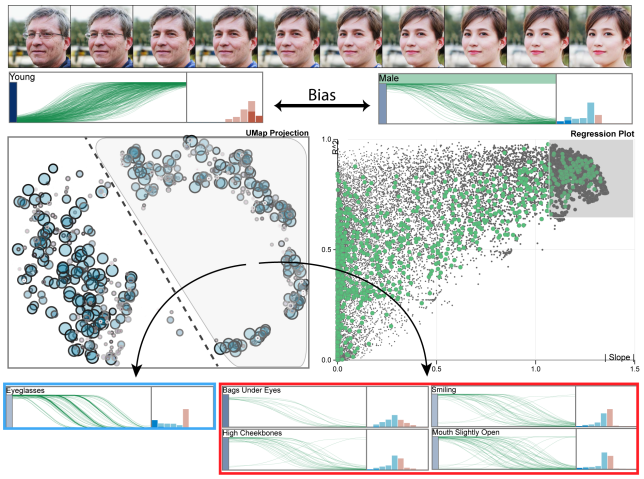

Figure 11: *Bias in the "Young" direction. The "Male" attribute's negative relationship with the "Young" attribute is the least explainable bias, relative to other correlations such as less "Gray hair" or less "Bags under eyes" which are attributes that we can easily expect to see from younger people in general. Additionally, in this example, we identify an interesting group of latent codes (right rim vs left rim indicated by a dotted line), which is divided by some artifacts around the eye and mouth area. Interaction available in* https:observablehq.com/d/6e85931d8e12c156.

In Figure 12 we show the Pale Skin direction, along with a default highlighted attribute for "Smiling". We find an age and smiling bias here from studying attributes brushed in (A). However, unlike the previous example in figure 11 the bias is not strong and widespread. This suggests that the bias is contained to a small group of latent codes which can be seen in (B).

When the bias is caused by a small group of codes in the latent space, it is often interesting to see what attributes the biased samples have in common. This is because the space that these codes reside might have some interesting properties such as insensitivity to certain factors of variation. A staged investigation can help researchers discover common attributes of the biased samples. In Figure 12, we can observe a cluster of biased samples found in the UMAP projection (B) - latent codes whose smiling becomes less as their skin color lightens. Common attributes $(A \cap B)$ of the walks in the cluster can be studied in the attribute view. In our example, most people who smile less are people who do not get younger. Also, assuming from the lack of consistency in "Male" attribute along with "Heavy Makeup" and "Wearing Lipstick" attributes, the codes in the cluster are likely to be females. Overall, we can conclude that this cluster contains codes of younger people (boys and girls whose gender is difficult to tell) who smile less.

7. Discussion

In this work we have shown how our visualization can enable users to understand the latent space of generative models from different angles. It can be used to discover directions and determine direction saliency, in turn assessing the distinctiveness of a direction.

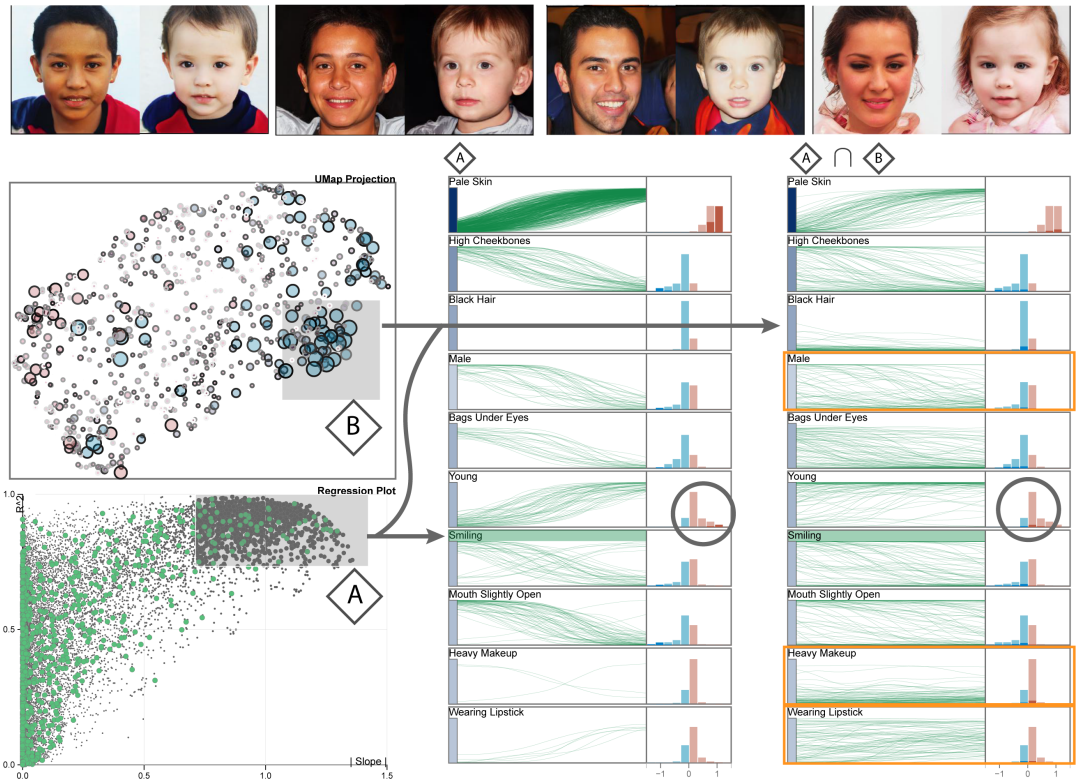

Figure 12: *Bias in the "Pale Skin" direction. From the set of codes selected in* (A)*, we can see that there exists a bias in the pale skin direction that tend to generate people who smiles less. One would normally not expect smiling to be associated with a skin tone. However, this bias is not widespread like in figure 11, but is present in a cluster of latent codes. The presence of an isolated bias calls for a deeper investigation as the cluster's common attributes can help researchers explain itself. The common attributes can be observed from an intersection of codes* $(A \cap B)$*. We can observe that the brushed latent codes in UMAP projection generate walks that are insensitive to the age direction (always young) while smiling less. Interactive example available in* https:observablehq.com/d/96f7594e43466ca9

It can also be used to determine if multiple attributes are consistently related, and if there exists correlation in the related attributes. Last, informativeness of a direction can easily be assessed through interactions that let users quickly reason about trends across attributes. The visualization can also be used to compare different directions, permitting the understanding of how different walking methods treat similar directions.

Beyond studying well-defined disentaglement properties, we showed that our visualization can be used to compare different walking methods, for which researchers studying direction finding methods may benefit from. Studying bias in the latent space is another use-case of our visualization, which benefits data scientists and general GAN community. Finally, we take a step further into the generator representation by linking projected convolutional activations with namable attributes.

One limitation of our research is that the quality of the latent space study hinges on the quality of attribute classifiers. Studying various aspects of a direction would be difficult without semantic attribute-based classifiers. We plan on studying how to use activations in the generator and discriminator to free the dependency on attributes. The second limitation comes from the fact that our visualization is primarily built to understand walks given a single direction. Although not impossible, it is more difficult to compare multiple directions at the same time. The ability to compare different directions is an important step in assessing disentanglement more broadly, e.g. ensuring a set of directions are distinct, yet completely describe the generative model.

Moving forward, we plan to refine our visualization to enable a comparison between multiple directions, as opposed to multiple attributes. This lets researchers understand a single latent code in-depth to reveal whether the latent code might fail to respond to directions, or has a strong property that cannot be altered via walking along directions. We believe such an interface can become useful for the purposes of image editing and content creation.

8. Acknowledgement

This work is supported in part by the National Science Foundation under grant number IIS-2007444, as well as under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. Released under LLNL-CONF-832727.

References

[AKBR19] ARNOUT H., KEHRER J., BRONNER J., RUNKLER T.: Visual evaluation of generative adversarial networks for time series data. *arXiv preprint arXiv:2001.00062* (2019). 3

[BDS18] BROCK A., DONAHUE J., SIMONYAN K.: Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096* (2018). 1

[CDH*16] CHEN X., DUAN Y., HOUTHOOFT R., SCHULMAN J., SUTSKEVER I., ABBEEL P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (2016), pp. 2180–2188. 3

[CL18] CHOO J., LIU S.: Visual analytics for explainable deep learning. *IEEE computer graphics and applications 38*, 4 (2018), 84–92. 1

[DBH18] DOŠILOVIĆ F. K., BRČIĆ M., HLUPIĆ N.: Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)* (2018), IEEE, pp. 0210–0215. 1

[EW18] EASTWOOD C., WILLIAMS C. K.: A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations* (2018). 1, 3

[FADK*18] FRID-ADAR M., DIAMANT I., KLANG E., AMITAI M., GOLDBERGER J., GREENSPAN H.: Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing 321* (2018), 321–331. 1

[GAOI19] GOETSCHALCKX L., ANDONIAN A., OLIVA A., ISOLA P.: Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 5744–5753. 1, 2

[GPAM*14] GOODFELLOW I. J., POUGET-ABADIE J., MIRZA M., XU B., WARDE-FARLEY D., OZAIR S., COURVILLE A., BENGIO Y.: Generative adversarial networks. *arXiv preprint arXiv:1406.2661* (2014). 1

[HAP*18] HIGGINS I., AMOS D., PFAU D., RACANIERE S., MATTHEY L., REZENDE D., LERCHNER A.: Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230* (2018). 3

[HHLP20] HÄRKÖNEN E., HERTZMANN A., LEHTINEN J., PARIS S.: Ganspace: Discovering interpretable gan controls. *arXiv preprint arXiv:2004.02546* (2020). 1, 3

[HKMG20] HEIMERL F., KRALJ C., MOLLER T., GLEICHER M.: embcomp: Visual interactive comparison of vector embeddings. *IEEE Transactions on Visualization and Computer Graphics* (2020). 3

[HMP*16] HIGGINS I., MATTHEY L., PAL A., BURGESS C., GLOROT X., BOTVINICK M., MOHAMED S., LERCHNER A.: beta-vae: Learning basic visual concepts with a constrained variational framework. 1, 3

[JCI19] JAHANIAN A., CHAI L., ISOLA P.: On the" steerability" of generative adversarial networks. *arXiv preprint arXiv:1907.07171* (2019). 2

[Jeb12] JEBARA T.: *Machine learning: discriminative and generative*, vol. 755. Springer Science & Business Media, 2012. 1

[KALL17] KARRAS T., AILA T., LAINE S., LEHTINEN J.: Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196* (2017). 1

[KLA19] KARRAS T., LAINE S., AILA T.: A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 4401–4410. 1, 2, 4, 6

[KTC*18] KAHNG M., THORAT N., CHAU D. H., VIÉGAS F. B., WATTENBERG M.: Gan lab: Understanding complex deep generative models using interactive visual experimentation. *IEEE transactions on visualization and computer graphics 25*, 1 (2018), 310–320. 3

[KWG*18] KIM B., WATTENBERG M., GILMER J., CAI C., WEXLER J., VIEGAS F., ET AL.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning* (2018), PMLR, pp. 2668–2677. 2

[LAR*19] LOCATELLO F., ABBATI G., RAINFORTH T., BAUER S., SCHÖLKOPF B., BACHEM O.: On the fairness of disentangled representations. *arXiv preprint arXiv:1905.13662* (2019). 3

[LJLH19] LIU Y., JUN E., LI Q., HEER J.: Latent space cartography: Visual analysis of vector space embeddings. In *Computer Graphics Forum* (2019), vol. 38, Wiley Online Library, pp. 67–78. 3

[LLWT15] LIU Z., LUO P., WANG X., TANG X.: Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision* (2015), pp. 3730–3738. 6

[LRT*14] LAFFONT P.-Y., REN Z., TAO X., QIAN C., HAYS J.: Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on Graphics (proceedings of SIGGRAPH) 33*, 4 (2014). 2

[LTH*17] LEDIG C., THEIS L., HUSZAR F., CABALLERO J., CUNNINGHAM A., ACOSTA A., AITKEN A., TEJANI A., TOTZ J., WANG Z., SHI W.: Photo-realistic single image super-resolution using a generative adversarial network, 2017. arXiv:1609.04802. 1

[MHM18] MCINNES L., HEALY J., MELVILLE J.: Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018). 5

[Mog16] MOGREN O.: C-rnn-gan: Continuous recurrent neural networks with adversarial training. *arXiv preprint arXiv:1611.09904* (2016). 1

[MSC*13] MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G., DEAN J.: Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546* (2013). 2

[MSI*18] MARIANI G., SCHEIDEGGER F., ISTRATE R., BEKAS C., MALOSSI C.: Bagan: Data augmentation with balancing gan. *arXiv preprint arXiv:1803.09655* (2018). 1

[PvdWR6] PERARNAU G., VAN DE WEIJER J., RADUCANU B., ÁLVAREZ J. M.: Invertible conditional gans for image editing. arXiv:1611.06355. 1

[RMC15] RADFORD A., METZ L., CHINTALA S.: Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015). 2

[SEBM20] SPINGARN-ELIEZER N., BANNER R., MICHAELI T.: Gan steerability without optimization. *arXiv preprint arXiv:2012.05328* (2020). 1

[SYTZ20] SHEN Y., YANG C., TANG X., ZHOU B.: Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020). 1, 2, 7

[SZ20] SHEN Y., ZHOU B.: Closed-form factorization of latent semantics in gans. *arXiv preprint arXiv:2007.06600* (2020). 1, 2, 3, 6

[TCAT17] TAN W. R., CHAN C. S., AGUIRRE H. E., TANAKA K.: Artgan: Artwork synthesis with conditional categorical gans. In *2017 IEEE International Conference on Image Processing (ICIP)* (2017), IEEE, pp. 3760–3764. 1

[VB20] VOYNOV A., BABENKO A.: Unsupervised discovery of interpretable directions in the gan latent space. In *International Conference on Machine Learning* (2020), PMLR, pp. 9786–9796. 1, 3

[YSZ21] YANG C., SHEN Y., ZHOU B.: Semantic hierarchy emerges in deep generative representations for scene synthesis. *International Journal of Computer Vision 129*, 5 (2021), 1451–1466. 1

[ZLK*17] ZHOU B., LAPEDRIZA A., KHOSLA A., OLIVA A., TORRALBA A.: Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017). 2

[ZXL*17] ZHANG H., XU T., LI H., ZHANG S., WANG X., HUANG X., METAXAS D.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks, 2017. arXiv:1612.03242. 1