

# Tensor Component Analysis for Interpreting the Latent Space of GANs

James Oldfield<sup>1</sup>

j.a.oldfield@qmul.ac.uk

Markos Georgopoulos<sup>2</sup>

m.georgopoulos@imperial.ac.uk

Yannis Panagakis<sup>3</sup>

yannisp@di.uoa.gr

Mihalis A. Nicolaou<sup>4</sup>

m.nicolaou@cyi.ac.cy

Ioannis Patras<sup>1</sup>

i.patras@qmul.ac.uk

<sup>1</sup> Queen Mary University of London, UK

<sup>2</sup> Imperial College London, UK

<sup>3</sup> University of Athens, Greece

<sup>4</sup> The Cyprus Institute, Cyprus

## Abstract

This paper addresses the problem of finding interpretable directions in the latent space of pre-trained Generative Adversarial Networks (GANs) to facilitate controllable image synthesis. Such interpretable directions correspond to transformations that can affect both the style and geometry of the synthetic images. However, existing approaches that utilise linear techniques to find these transformations often fail to provide an intuitive way to separate these two sources of variation. To address this, we propose to a) perform a *multilinear* decomposition of the tensor of intermediate representations, and b) use a tensor-based regression to map directions found using this decomposition to the latent space. Our scheme allows for both linear edits corresponding to the individual modes of the tensor, and non-linear ones that model the multiplicative interactions between them. We show experimentally that we can utilise the former to better separate style- from geometry-based transformations, and the latter to generate an extended set of possible transformations in comparison to prior works. We demonstrate our approach's efficacy both quantitatively and qualitatively compared to the current state-of-the-art.

## 1 Introduction

Over the past few years, GANs [8] have continued to push the state-of-the-art forward for the task of image synthesis. Many recent works have proposed more sophisticated architectures for improving the quality of the generated images: such as with the use of transposed convolutions [29], progressive growing [16], or with explicit modulation of the style content [10]. As the quality of the images generated by these methods continues to improve, there is increasing interest in exploring how one can better control the image synthesis process. A prominent research direction with this goal in mind is that of finding directions in the latent space that reliably affect interpretable modes of variation [8, 12] in the generated

Transformation type	‘Style’	‘Geometry’	‘Multilinear Mixing’
Examples	illumination, colour, background	yaw, pitch, translation	deformation, skew, distort
Prominent modes	Channel mode	Spatial modes	Interactions between modes

Table 1: The hierarchy of transformations obtainable (and where) with our method.

images [30, 31, 32, 33, 34, 35, 36]. Prior works showed that with supervision, one can isolate factors such as age, gender, and pose [31, 32]—facilitating the ability to modify these attributes in an image by a desired amount. Despite this success, the necessary supervision requires expensive manual labour, highlighting the need for an unsupervised discovery of such directions. In this vein, recent methods use auxiliary networks to search for ‘diverse’ image transformations [32, 33], decompose the weights defining the mapping between layers [30], or decompose the intermediate generator’s representations directly [32]. However, the latter’s approach of treating this tensor of intermediate activations as a vector entangles the variation in both the spatial and channel modes. With these approaches, there is often no intuitive way to separate different types of transformations. Such an ability has recently been shown to be useful for a number of downstream tasks, such as saliency detection [36] and unsupervised object segmentation [37]. To this end, we take inspiration from the categorisations introduced in [38] and focus on locating two different types of interpretable directions: *style*-based directions (such as illumination and colour) and *geometry*-based directions (such as orientation and translation). This categorisation is defined in Table 1, along with examples of the types of transformations we seek to find.

Motivated by findings from the style transfer literature [32], we suggest in this work that the modes of the tensors of activations in a generator can be useful for isolating these different types of semantic transformations. To address this, we propose a multilinear approach that finds interpretable directions in the latent activations’ natural tensorial form  $\mathcal{Z} \in \mathbb{R}^{C \times H \times W}$  as shown in Fig. 1. We learn a separate basis  $\mathbf{U}^{(C)}, \mathbf{U}^{(H)}, \mathbf{U}^{(W)}$  for the channel, height, and width modes of the tensors, respectively, to locate the various types of transformations described above. Such a multilinear treatment comes with a number of benefits beyond being computationally cheaper than its linear counterpart. Firstly, this leads to an implicit separation of style and geometry: this allows one to find interpretable directions corresponding semantically to each mode of the tensor. What’s more, we show how the multiplicative interactions [32] of basis vectors across the modes of the tensor (‘multilinear mixing’ in Table 1) correspond to transformations unobtainable with the linear treatment (such as forehead shape, where both the width and height dimensions influence the attribute together in a non-uniform manner). We propose the use of a tensor regression to map these combinations of basis vectors in the form of a tensor back to latent space—with a low-rank structure to offer flexibility over the extent to which the interpretable directions are similar to the transformations found in the training data. We demonstrate with a series of experiments on multiple generator architectures and datasets the validity of our method—including showing superior disentanglement both qualitatively and quantitatively over prior work. Our main contributions can be summarised as follows:

- We propose an intuitive way to separate different types of interpretable directions in a GAN by decomposing the activation maps in the generator in a multilinear fashion. We show how the linear approach of [32] can be framed as a special case.
- We show that by modelling the *interactions* of basis vectors across modes, one can recover transformations that are the influence of multiple modes in a non-uniform

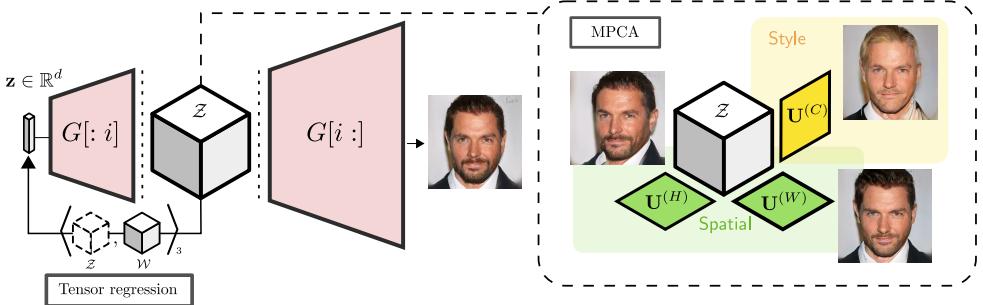


Figure 1: An overview of our proposed method: on a pre-trained generator’s intermediate features  $Z \in \mathbb{R}^{C \times H \times W}$  we perform multilinear PCA (right) to learn a basis for each mode of the tensor-localising different types of interpretable directions to each mode. We learn a tensor regression from the activations back to the latent code, allowing us to then map combinations of the multilinear bases back to interpretable directions in latent space.

fashion (such as thickening of the forehead), and are hence unobtainable with linear approaches.

- We propose the use of a tensor regression for mapping activations back to their latent code, which provides regularisation in the form of a low-rank structure on the weights.
- We demonstrate both qualitatively and quantitatively that prominent directions along each mode learnt with our method correspond to attributes such as hair colour, pitch, or yaw. These found directions showcase superior disentanglement over the state-of-the-art methods in terms of the style or geometry categories of transformations identified in Table 1.

## 2 Related work

**GANs & Interpretable directions** The seminal work of GANs [8] showed how, using adversarial training, one can train a so-called generator to map a latent noise vector to a high resolution synthetic image. This generator is often realised with a convolutional neural network [16, 29], with various tricks having been developed to improve the quality of the samples [4, 10, 18]. Interestingly, [29] showed that one can perform arithmetic in the latent space that affects predictable changes in image space. Since these works, a host of methods have been proposed to explore the latent structure in these generators by imposing structure at training-time [4, 24] or more recently in the pre-trained generators themselves [1, 10, 31, 32, 34, 36]. However, of the approaches that decompose the intermediate features directly (such as [10]), a linear decomposition is applied—where we argue a multilinear one can be more suitable in providing an ability to locate different categories of transformation.

**Tensor methods for visual data** The use of multilinear structures or decompositions to represent and interpret visual data has a rich history. For example, bilinear models have been used to separate style and content [33], and multilinear ones for learning the structure of visual data more generally [35, 38, 39]. More recently, a popular approach is to combine these tensor methods with deep learning methodologies [2]. For example, for modelling [8, 10]

multiplicative interactions [15], disentangling the modes of variation [20], or for interpreting or decomposing the weights and operations of a deep neural network [8, 9, 23].

## 3 Methodology

In this section, we describe our method in detail. We first provide an overview of the relevant notation and definitions in Section 3.1. We follow in Section 3.2 by formulating the problem of learning interpretable directions, including the PCA solution in Section 3.2.1. In Section 3.3, we introduce the proposed multilinear approach (formulating GANSpace as a special case), and finally in Section 3.4 we detail how we use these learnt directions to modify the latent code.

### 3.1 Notation

Firstly, we detail the notation we adopt throughout this paper, and provide definitions of the relevant operations<sup>1</sup>. We use uppercase (lowercase) boldface letters to denote matrices (vectors), e.g.  $\mathbf{X}$  ( $\mathbf{x}$ ), and calligraphic letters for tensors, e.g.  $\mathcal{X}$ . We use  $\mathbf{1}_d \in \mathbb{R}^d$  to denote the vector of ones from the main diagonal of a  $d$ -dimensional identity matrix—the subscript of which we sometimes omit for brevity. The **Kronecker product** of two matrices  $\mathbf{X} \in \mathbb{R}^{I_1 \times I_2}$  and  $\mathbf{Y} \in \mathbb{R}^{J_1 \times J_2}$  is defined as

$$\mathbf{X} \otimes \mathbf{Y} = \begin{bmatrix} x_{11} \mathbf{Y} & \cdots & x_{1I_2} \mathbf{Y} \\ \vdots & \ddots & \vdots \\ x_{I_1 1} \mathbf{Y} & \cdots & x_{I_1 I_2} \mathbf{Y} \end{bmatrix} \in \mathbb{R}^{I_1 \cdot J_1 \times I_2 \cdot J_2}.$$

We refer to each element of an  $N^{th}$  order tensor  $\mathcal{X}$  using  $N$  indices, i.e.,  $\mathcal{X}(i_1, i_2, \dots, i_N) \doteq x_{i_1 i_2 \dots i_N} \in \mathbb{R}$ . The **mode- $n$  fibers** of a tensor are the column vectors formed when fixing all but one of the indices (e.g.  $\mathbf{x}_{:jk}$ ), and can be seen as a higher-order analogue of matrices' rows and columns. For a tensor  $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$ , we can arrange its mode- $n$  fibers along the columns a matrix, giving us the notion of the **mode- $n$  unfolding** denoted as  $\mathbf{X}_{(n)} \in \mathbb{R}^{I_n \times \bar{I}_n}$  with  $\bar{I}_n = \prod_{t=1, t \neq n}^N I_t$ . Lastly, the **mode- $n$  (matrix) product** of a tensor  $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$  and a matrix  $\mathbf{W} \in \mathbb{R}^{J \times I_n}$  is denoted by  $\mathcal{X} \times_n \mathbf{W} \in \mathbb{R}^{I_1 \times \dots \times I_{n-1} \times J \times I_{n+1} \times \dots \times I_N}$ . Most usefully for this paper, the mode- $n$  matrix product can be expressed in terms of the unfolded tensors as

$$\mathcal{Y} = \mathcal{X} \times_n \mathbf{W} \quad \Leftrightarrow \quad \mathbf{Y}_{(n)} = \mathbf{W} \mathbf{X}_{(n)}.$$

### 3.2 Problem formulation

A pretrained generator  $G$  receives a low-dimensional vector latent code  $\mathbf{z} \in \mathbb{R}^d$  and maps it to a synthetic image  $G(\mathbf{z}) = \mathcal{X} \in \mathbb{R}^{C \times H \times W}$ .  $G$  is usually implemented as a series of convolutional layers, meaning that the image has a number of intermediate representations  $\mathcal{Z}$  in the form of a collection of feature maps. In traditional generator architectures, the output image  $\mathcal{X}$  is parameterised fully by its latent code  $\mathbf{z}$ . Therefore by modifying the latent code, one can affect changes in the resulting synthetic image. The goal of learning interpretable directions is that of finding a latent vector  $\mathbf{z}'$  corresponding to a high-level attribute of interest (such as

<sup>1</sup>We refer readers to [20] for a more detailed overview.

‘hair color’) in image space. One can then modify the original latent code by some amount  $\alpha \in \mathbb{R}$  to generate the modified image  $\mathcal{X}' = G(\mathbf{z} + \alpha \cdot \mathbf{z}')$ .

### 3.2.1 The linear solution: GANSpace

Let  $\{\mathcal{Z}_1, \mathcal{Z}_2, \dots, \mathcal{Z}_M\}$  be a batch of  $M$  feature maps from an intermediate layer in a pretrained generator, each of which is a third order tensor  $\mathcal{Z}_m \in \mathbb{R}^{C \times H \times W}$  with channel, height, and width modes. The PCA-based method of GANSpace [12] finds interpretable directions in the traditional architectures by first vectorising these tensors  $\text{vec}(\mathcal{Z}_m)$  and then learning a PCA basis  $\mathbf{U}$  that admits the following decomposition

$$\text{vec}(\mathcal{Z}_m) = \mathbf{U}\mathbf{U}^\top \text{vec}(\mathcal{Z}_m) \quad (1)$$

$$= \text{vec}(\mathcal{Z}_m) \times_1 \mathbf{U}\mathbf{U}^\top. \quad (2)$$

This basis  $\mathbf{U}$  is then regressed back to the latent space with linear transformation  $\mathbf{W}$ , the columns of which then contain the interpretable directions in the original latent space. One can then make edits with  $\mathcal{X}' = G(\mathbf{z} + \mathbf{Ws})$ , where  $\mathbf{s}$  is a ‘selector’ vector that takes a linear combination of the desired columns of the basis.

## 3.3 Learning multilinear directions

Rather than living in a single  $C \cdot H \cdot W$ -dimensional vector space however, each activation tensor  $\mathcal{Z}_m$  can be seen more naturally as belonging to a tensor space formed by the outer product of *three* vector spaces  $\mathbb{R}^C \otimes \mathbb{R}^H \otimes \mathbb{R}^W$ . We thus retain the tensorial structure and perform a *multilinear* PCA [10], allowing us to write each of the  $M$  tensors in the batch as

$$\mathcal{Z}_m = \mathcal{Z}_m \times_1 \mathbf{U}^{(C)} \mathbf{U}^{(C)\top} \times_2 \mathbf{U}^{(H)} \mathbf{U}^{(H)\top} \times_3 \mathbf{U}^{(W)} \mathbf{U}^{(W)\top}, \quad (3)$$

where each  $\mathbf{U}^{(n)}$  forms an orthonormal basis for the space spanned by all  $M$  tensors’ mode- $n$  fibers. From this it is clear that the GANSpace [12] approach as we formulate it using the mode-1 product in Eq. (2) is a special case of our proposed formulation in Eq. (3)—instead operating on the flattened tensor with a single factor matrix. To further shed light on the connection between our formulation and the linear approach of GANSpace in Eq. (1), we can rewrite each activation tensor in Eq. (3) in its vectorised form [10] as

$$\text{vec}(\mathcal{Z}_m) = \left( \mathbf{U}^{(W)} \mathbf{U}^{(W)\top} \otimes \mathbf{U}^{(H)} \mathbf{U}^{(H)\top} \otimes \mathbf{U}^{(C)} \mathbf{U}^{(C)\top} \right) \text{vec}(\mathcal{Z}_m), \quad (4)$$

highlighting the crucial differences between the linear and multilinear methods: MPCA models the interactions of the columns of all three bases.

### 3.3.1 Computing the multilinear basis

The factor matrices  $\mathbf{U}^{(n)\top}$  in Eq. (3) project the input tensor to a tensor subspace where the total scatter is maximised [2, 10]—this objective is a higher-order analogue of that of regular PCA. We thus follow Section. III. B of Lu et al. [10] and compute factor matrix  $\mathbf{U}^{(n)}$  as the matrix of left-singular vectors from the SVD of the following mode- $n$  total scatter matrix

$$\mathbf{U}^{(n)} \Sigma \mathbf{V}^{(n)\top} = \sum_{m=1}^M (\mathbf{Z}_{m(n)} - \bar{\mathbf{Z}}_{(n)}) (\mathbf{Z}_{m(n)} - \bar{\mathbf{Z}}_{(n)})^\top, \quad (5)$$

where  $\mathbf{Z}_{m(n)}$  is the mode- $n$  unfolding of the  $m^{\text{th}}$  GAN sample’s activation tensor, and  $\bar{\mathbf{Z}}_{(n)}$  is the mean of all mode- $n$  unfoldings. The columns of each of these factor matrices contain the principal directions for each mode. Pseudocode outlining this procedure in a pre-trained GAN can be found in the supplementary material.

### 3.4 Editing with multilinear directions

Rather than taking a linear combination of the basis vectors like in the linear setting of Section 3.2.1, we propose to instead form what we call an ‘edit tensor’,  $\mathcal{Z}' \in \mathbb{R}^{C \times H \times W}$  as a combination of the basis vectors across the modes, either separately or together. These different types of combinations of the basis vectors produce the various transformations in Table 1. A graphical illustration of this process can be found in the supplementary material.

**Mode-wise edits** Each column of  $\mathbf{U}^{(n)}$  is a basis vector for the space spanned by the mode- $n$  fibers. We perform a mode- $n$  edit by adding one or more of these basis vectors to all mode- $n$  fibers of  $\mathcal{Z}'$ . For example, to perform an edit using the  $i^{\text{th}}$  channel basis vector we add  $\alpha_i \cdot \mathbf{u}_i^{(C)} \circ \mathbf{1}_H \circ \mathbf{1}_W$  to the edit tensor—where  $\alpha_i$  is a scalar controlling the relative weight. More generally, to add the  $i^{\text{th}}$  basis vector for mode  $n$ , we add the term  $\mathcal{S}_n \times_n \mathbf{U}^{(n)}$  to the edit tensor, where all elements of the mode- $n$  unfolding of  $\mathcal{S}_n$  are zero except for its  $i^{\text{th}}$  row that is set to  $\mathbf{S}_{n(n)}(i, :) := \alpha_i \cdot \mathbf{1}$ . We will experimentally show in Section 4.1 that, broadly speaking, edits along the channel modes generate changes to the style of the image, whilst the major geometric changes can be generated by basis vectors for the spatial modes.

**Multilinear mixing** In addition to making edits along a single mode, we can alternatively model the multiplicative interactions [■] of basis vectors *between* modes—we call this ‘multilinear mixing’. To model the third-order interactions of the  $i^{\text{th}}$ ,  $j^{\text{th}}$ , and  $k^{\text{th}}$  basis vectors for the three modes respectively, we define a selector tensor  $\mathcal{S}_{CHW} \in \mathbb{R}^{C \times H \times W}$  with  $\mathcal{S}_{CHW}(i, j, k) := \alpha_{ijk}$ , and set the edit tensor as  $\mathcal{S}_{CHW} \times_1 \mathbf{U}^{(C)} \times_2 \mathbf{U}^{(H)} \times_3 \mathbf{U}^{(W)}$ , that is, a rank-1 tensor formed as the outer product  $\alpha_{ijk} \cdot \mathbf{u}_i^{(C)} \circ \mathbf{u}_j^{(H)} \circ \mathbf{u}_k^{(W)}$ . In the general case when multiple directions are manipulated simultaneously this term can be written as a sum over rank-1 tensors formed by outer products of the desired vectors from each basis, weighted by the elements in  $\mathcal{S}_{CHW}$ . Second-order terms are obtained analogously as a rank-1 matrix which is replicated along each slice of the edit tensor with an outer product with the  $\mathbf{1}$ -vector. Combining the 1st-, 2nd-, and 3rd-order terms leads to the most general form of the edit tensor as

$$\begin{aligned} \mathcal{Z}' = & \mathcal{S}_C \times_1 \mathbf{U}^{(C)} + \mathcal{S}_H \times_2 \mathbf{U}^{(H)} + \mathcal{S}_W \times_3 \mathbf{U}^{(W)} \\ & + \mathcal{S}_{CH} \times_1 \mathbf{U}^{(C)} \times_2 \mathbf{U}^{(H)} + \dots + \mathcal{S}_{CHW} \times_1 \mathbf{U}^{(C)} \times_2 \mathbf{U}^{(H)} \times_3 \mathbf{U}^{(W)}. \end{aligned} \quad (6)$$

#### 3.4.1 Mapping edits to latent space

To use these directions found in the activation space to edit the original latent code, we transform this edit tensor back to the original latent space [■]. Concretely, we seek to learn a mapping from the original activation tensor  $\mathcal{Z} \in \mathbb{R}^{C \times H \times W}$  to its corresponding latent code  $\mathbf{z} \in \mathbb{R}^d$ . Using this mapping, we can then generate the latent code  $\mathbf{z}'$  for a desired edit tensor

$\mathcal{Z}'$ . To solve this, we propose a tensor regression [10, 22] of the form

$$\mathbf{z} = \langle \mathcal{Z}, \mathcal{W} \rangle_3 = \sum_c \sum_h \sum_w \mathcal{Z}(c, h, w) \mathcal{W}(c, h, w, :), \quad (7)$$

with weight tensor  $\mathcal{W} \in \mathbb{R}^{C \times H \times W \times d}$ . That is to say, we take a “generalised inner product” [22] along the last 3 modes of  $\mathcal{Z}$  and first 3 modes of  $\mathcal{W}$ . This gives us the objective function for the regression task as

$$\mathcal{L}_{rec} = \mathbb{E}_{\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)} \left[ \|\mathbf{z} - \langle \mathcal{Z}, \mathcal{W} \rangle_3\|_2^2 \right] + \lambda \|\mathcal{W}\|_2^2. \quad (8)$$

We explore imposing a low-rank Tucker structure on  $\mathcal{W}$  in order to reduce the number of parameters and perform regularisation, training the factors of the Tucker decomposition forming  $\mathcal{W}$  in Eq. (8) with gradient descent using TensorLy [21]. Finally, following the linear case in Section 3.2.1, we perform the edit with  $\mathcal{X}' = G(\mathbf{z} + \langle \mathcal{Z}', \mathcal{W} \rangle_3)$ , where  $\mathcal{Z}'$  is given by Eq. (6).

## 4 Experiments

In this section, we present a series of experiments to validate the proposed method. We first showcase in Section 4.1 how we can find directions along each mode that correspond intuitively to either geometry- or style-based attributes, along with the multilinear mixes. Finally, we present quantitative results in Section 4.2 that show superior disentanglement compared to the state-of-the-art.



Figure 2: Edits performed along the spatial and channel modes separately, in a variety of generators and datasets. For these experiments, we use a low-rank Tucker decomposition for the regression tensor.

**Implementation details** We experiment on both the traditional (transpose) convolutional architecture of **ProgGAN** and the style-based generator from **StyleGAN**. We apply the multilinear decomposition after the first block of convolutions for both networks, and map these activations back to their corresponding latent codes (or style vectors, for StyleGAN). The images generated by these networks are not always high quality however, and therefore in this section we manually select initial seeds that produce realistic images to showcase our results on.

## 4.1 Qualitative results



Figure 3: Walking along a single found direction in the channel basis by the same amount for 8 random seed images, we find each image is transformed in the same manner.

**Mode-wise edits** In this section, we demonstrate qualitatively some of our method’s learnt directions for each mode. We find that by making edits along the channels of the activation maps we tend to affect changes to the style of the image, leaving the geometry of the image (such as its pose) largely untouched. This is demonstrated qualitatively in Fig. 2—for example, the channel basis vectors affect semantic changes such as the colour of a person’s hair or skin. We show how such directions affect different images in a consistent manner in Fig. 3. On the other hand, we frequently find at least one of the major geometry-based directions in the relevant spatial mode. For example, for CelebA-HQ [10] on ProgGAN, we find ‘pitch’ in the basis for the horizontal mode, and ‘yaw’ in the vertical basis. As with all PCA-based methods however, the number of semantically relevant directions we can find in each mode is limited by the variation in the original training data.

**Multilinear mixing** As detailed in Section 3.4, our method can alternatively model the interactions of the basis vectors. Here we show experimentally that such ‘multilinear mixing’ can generate a unique set of transformations for which the influence of more than one mode is necessary. In particular, we list examples of these transformations we observe to be possible experimentally in the right-most column of Table 1. We show in Fig. 4 some of these on ProgGAN, allowing us fine-grained control over attributes such as forehead size and face shape. We find this ability to affect these kinds of coarse changes to not be possible with the baseline methods—highlighting the importance of explicitly treating the modes separately and modelling the interactions between their bases. We find a high-rank Tucker structure on the weight tensor is beneficial to best capture these coarse changes in the higher-order terms.



Figure 4: Edits found in the third-order interactions of bases for ProgGAN (with a rank-256, 4, 4, 512 Tucker decomposition on the weight tensor).

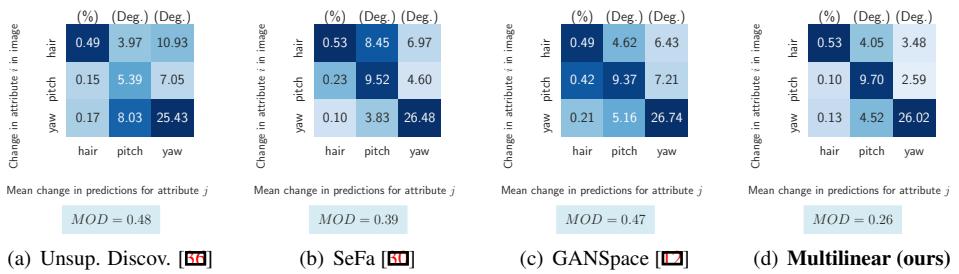

MOD = 0.48
MOD = 0.39
MOD = 0.47
MOD = 0.26

Figure 5: (a) Mean difference in predictions (columns) between the original images and their edited versions for the target attribute (rows), on various baselines. The mean of the (column-normalised) off-diagonals ( $MOD \downarrow$ ) is shown below each.

## 4.2 Quantitative results

Finally, we quantify how well our method can recover implicitly disentangled directions corresponding to style and geometry. We use the same model trained on ProgGAN with which we generated Fig. 2, manually identifying a prominent recovered direction for each mode: hair colour (channels), yaw (vertical), and pitch (horizontal). We synthesise 100 random images, and 3 edited versions of each image, walking along the direction corresponding to each of the three attributes. We obtain predictions for these three attributes on all 400 images using Microsoft Azure<sup>2</sup>. Finally, we compute the mean absolute difference between the predictions for each attribute on the unedited images and on the 3 edited versions, to see how changing one attribute affects the predictions for the other two. We collect these values in a matrix  $\mathbf{A} \in \mathbb{R}^{N \times N}$  (shown for various baselines in Fig. 5 with  $N := 3$ ), with  $a_{i,j}$  being the mean difference in predictions for attribute  $i$  between the raw images and those with attribute  $j$  edited. It's clear from the small values in the off-diagonals of Fig. 5(d), that these directions affect only the single target factor of variation to a greater extent than the baseline results in Figs. 5(a) to 5(c). We can quantify this by taking the mean of the off-diagonals ( $MOD$ ) for each method with  $MOD = \frac{1}{N^2 - N} \left( \sum_{i,j} \hat{a}_{i,j} - \text{tr}(\hat{\mathbf{A}}) \right)$ , where  $\hat{\mathbf{A}}$  is the column-normalised  $\mathbf{A}$ .

<sup>2</sup><https://azure.microsoft.com/en-gb/services/cognitive-services/face/>



Figure 6: Linearly interpolating along the basis vectors for ‘yaw’, ‘pitch’, and ‘blond hair’ for both our method and GANspace. As can be seen, the linear model’s ‘pitch’ direction is entangled with hair colour, whereas ours exhibits clearer visual disentanglement.

Our method achieves a notably lower score than all baselines. We see qualitatively in our GANSpace comparison in Fig. 6 that the ‘pitch’ attribute identified in GANspace clearly leaks style information to a greater extent than our direction for the ‘pitch’ attribute.

## 5 Conclusion

In this paper, we have presented a method that offers an intuitive way to find different types of interpretable transformations in a pre-trained GAN. We achieve this by decomposing the generator’s activations in a multilinear manner, and regressing back to the latent space. We showed that one can find directions along each mode of the tensor that correspond semantically to the relevant mode (for example, we find the ‘yaw’ direction in the ‘vertical’ basis). Additionally we showed that by modelling the multiplicative interactions of basis vectors across the modes, we recover an extended set of directions we find to be unobtainable with other methods in the literature. We showed that our found directions achieve superior disentanglement over prior work. In our experiments, we found the choice of rank on the weight tensor to play an important role in the quality of our found directions. In future work we plan to develop techniques to determine the appropriate rank automatically.

**Acknowledgements** This work was supported by a grant from The Cyprus Institute on Cyclone under project ID p055, and the EU H2020 AI4Media No. 951911 project.

## References

- [1] Terence Broad, Frederic Fol Leymarie, and Mick Grierson. Network bending: Expressive manipulation of deep generative models. In *Artificial Intelligence in Music, Sound, Art and Design*, pages 20–36. Springer International Publishing, 2021.
- [2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019.
- [3] Adrian Bulat, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. Incremental multi-domain learning with network latent tensor factorization. In *Proc. AAAI Conf. Artif. Intell. (AAAI)*, volume 34, pages 10470–10477, 2020.
- [4] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Proc. Adv. Neural Inform. Process. Syst. (NeurIPS)*, volume 29, 2016.
- [5] Grigoris G Chrysos, Stylianos Moschoglou, Giorgos Bouritsas, Yannis Panagakis, Jiankang Deng, and Stefanos Zafeiriou. P-nets: Deep polynomial neural networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 7325–7335, 2020.
- [6] Markos Georgopoulos, Grigoris Chrysos, Maja Pantic, and Yannis Panagakis. Multilinear latent conditioning for generating unseen attribute combinations. In *Proc. Int. Conf. Mach. Learn. (ICML)*, volume 119, pages 3442–3451, 2020.
- [7] Markos Georgopoulos, James Oldfield, Mihalis Nicolaou, Yannis Panagakis, and Maja Pantic. Mitigating demographic bias in facial datasets with style-based multi-attribute transfer. *Int. J. Comput. Vis. (IJCV)*, 2021.
- [8] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *Proc. Int. Conf. Comput. Vis. (ICCV)*, pages 5744–5753, 2019.
- [9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2014.
- [10] W. Guo, I. Kotsia, and I. Patras. Tensor learning for regression. *IEEE Trans. Image Process.*, 21(2):816–827, 2012.
- [11] Haiping Lu, K.N. Plataniotis, and A.N. Venetsanopoulos. MPCA: Multilinear Principal Component Analysis of Tensor Objects. *IEEE Trans. Neural Netw.*, 19:18–39, 2008.
- [12] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. GANSpace: Discovering interpretable GAN controls. In *Proc. Adv. Neural Inform. Process. Syst. (NeurIPS)*, volume 33, pages 9841–9850, 2020.
- [13] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proc. Int. Conf. Comput. Vis. (ICCV)*, pages 1501–1510, 2017.

- [14] Ali Jahanian, Lucy Chai, and Phillip Isola. On the "steerability" of generative adversarial networks. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.
- [15] Siddhant M Jayakumar, Wojciech M Czarnecki, Jacob Menick, Jonathan Schwarz, Jack Rae, Simon Osindero, Yee Whye Teh, Tim Harley, and Razvan Pascanu. Multiplicative interactions and where to find them. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.
- [16] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018.
- [17] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, June 2019.
- [18] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, June 2020.
- [19] Tamara G. Kolda. Multilinear operators for higher-order decompositions. Technical report, Sandia National Laboratories, April 2006.
- [20] Tamara G. Kolda and Brett W. Bader. Tensor Decompositions and Applications. *SIAM Review*, 51(3):455–500, August 2009.
- [21] Jean Kossaifi, Yannis Panagakis, Anima Anandkumar, and Maja Pantic. TensorLy: Tensor learning in Python. *J. Mach. Learn. Res.*, 20(26), 2019.
- [22] Jean Kossaifi, Zachary C Lipton, Arinbjörn Kolbeinsson, Aran Khanna, Tommaso Furlanello, and Anima Anandkumar. Tensor regression networks. *J. Mach. Learn. Res.*, 21:1–21, 2020.
- [23] Jean Kossaifi, Antoine Toisoul, Adrian Bulat, Yannis Panagakis, Timothy M Hospedales, and Maja Pantic. Factorized higher-order CNNs with an application to spatio-temporal emotion estimation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 6060–6069, 2020.
- [24] Wonkwang Lee, Donggyun Kim, Seunghoon Hong, and Honglak Lee. High-fidelity synthesis with disentangled representation. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pages 157–174, 2020.
- [25] Haiping Lu, Konstantinos Plataniotis, and Anastasios Venetsanopoulos. A survey of multilinear subspace learning for tensor data. *Pattern Recog.*, 44:1540–1551, 08 2011.
- [26] Haiping Lu, Konstantinos N. Plataniotis, and Anastasios Venetsanopoulos. *Multilinear Subspace Learning: Dimensionality Reduction of Multidimensional Data*. Chapman & Hall/CRC, 1st edition, 2013.
- [27] Yannis Panagakis, Jean Kossaifi, Grigorios G. Chrysos, James Oldfield, Mihalis A. Nicolaou, Anima Anandkumar, and Stefanos Zafeiriou. Tensor methods in computer vision and deep learning. *Proc. IEEE Proc.*, 109(5):863–890, 2021.

- [28] Antoine Plumerault, Hervé Le Borgne, and Céline Hudelot. Controlling generative models with continuous factors of variations. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.
- [29] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016.
- [30] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in GANs. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2021.
- [31] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of GANs for semantic face editing. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020.
- [32] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. InterFaceGAN: Interpreting the disentangled face representation learned by GANs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.
- [33] Joshua B. Tenenbaum and William T. Freeman. Separating Style and Content with Bilinear Models. *Neural Computation*, 12(6):1247–1283, June 2000.
- [34] Christos Tzelepis, Georgios Tzimiropoulos, and Ioannis Patras. WarpedGANSpace: Finding non-linear RBF paths in GAN latent space. In *Proc. Int. Conf. Comput. Vis. (ICCV)*, pages 6393–6402, October 2021.
- [35] M. Alex O. Vasilescu and Demetri Terzopoulos. Multilinear Analysis of Image Ensembles: TensorFaces. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, volume 2350, pages 447–460. Berlin, Heidelberg, 2002.
- [36] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the GAN latent space. In *Proc. Int. Conf. Mach. Learn. (ICML)*, pages 9786–9796, 2020.
- [37] Andrey Voynov, Stanislav Morozov, and Artem Babenko. Big GANs are watching you: Towards unsupervised object segmentation with off-the-shelf generative models. *arXiv preprint arXiv:2006.04988*, 2020.
- [38] Mengjiao Wang, Yannis Panagakis, Patrick Snape, and Stefanos Zafeiriou. Learning the Multilinear Structure of Visual Data. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 6053–6061, July 2017.
- [39] Mengjiao Wang, Yannis Panagakis, Patrick Snape, and Stefanos P Zafeiriou. Disentangling the modes of variation in unlabelled data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(11):2682–2695, 2017.
- [40] Mengjiao Wang, Zhixin Shu, Shiyang Cheng, Yannis Panagakis, Dimitris Samaras, and Stefanos Zafeiriou. An adversarial neuro-tensorial approach for learning disentangled representations. *Int. J. Comput. Vis. (IJCV)*, 127(6):743–762, 2019.