# SC$^2$GAN: Rethinking Entanglement by Self-correcting Correlated GAN Space

Zikun Chen[1], Han Zhao[2], Parham Aarabi[3], Ruowei Jiang[1]

[1]ModiFace Inc.
[2]University of Illinois at Urbana-Champaign
[3]University of Toronto

## Abstract

*Generative Adversarial Networks (GANs) can synthesize realistic images, with the learned latent space shown to encode rich semantic information with various interpretable directions. However, due to the unstructured nature of the learned latent space, it inherits the bias from the training data where specific groups of visual attributes that are not causally related tend to appear together, a phenomenon also known as spurious correlations, e.g., age and eyeglasses or women and lipsticks. Consequently, the learned distribution often lacks the proper modelling of the missing examples. The interpolation following editing directions for one attribute could result in entangled changes with other attributes. To address this problem, previous works typically adjust the learned directions to minimize the changes in other attributes, yet they still fail on strongly correlated features. In this work, we study the entanglement issue in both the training data and the learned latent space for the StyleGAN2-FFHQ model. We propose a novel framework SC$^2$GAN that achieves disentanglement by reprojecting low-density latent code samples in the original latent space and correcting the editing directions based on both the high-density and low-density regions. By leveraging the original meaningful directions and semantic region-specific layers, our framework interpolates the original latent codes to generate images with attribute combination that appears infrequently, then inverts these samples back to the original latent space. We apply our framework to pre-existing methods that learn meaningful latent directions and showcase its strong capability to disentangle the attributes with small amounts of low-density region samples added.*

## 1. Introduction

Recent advances of Generative Adversarial Networks (GANs) [10], such as StyleGAN [14, 15, 13] and BigGAN [7], boost a remarkable success for synthesizing photo-realistic images. In addition to a variety of real-world applications such as image-to-image translation [12, 27] or text-to-image translations [26, 16], another line of work [4, 5, 25, 11, 20, 8] that studies the interpretability of GANs has also generated increasing attention in the community. These works study the learned latent space by identifying semantically meaningful directions and interpolating along the learned directions. However, challenges remain to perfectly disentangle correlated features such as age and eyeglasses while obtaining valid feature controls, due to the biased learned distribution.

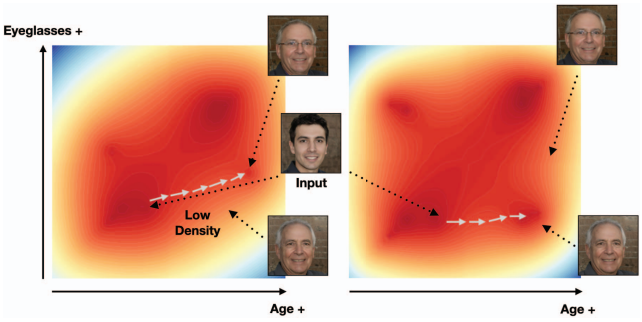



Figure 1: Illustration of our method: original latent distribution (left) and self-corrected latent distribution (right). The intensity of the color indicates the density of different subregions in the learned latent space. The white arrows show the interpolation directions. The "eyeglasses" direction and the "age" direction are more orthogonal to each other in the self-corrected latent space.

To tackle the challenge of entangled features, prior works have largely taken three approaches to obtain disentangled controls: orthogonalization of the learned directions [20, 23, 3], controls based on semantic masks [25], and gradient-based channel filtering [8]. Orthogonalization of the learned directions [20, 23, 3] follows the assumption that for any learned direction, a change along another or-

thogonal axis should not affect the feature for the learned direction. For example, if $A$ and $B$ are two orthogonal directions that define two hyperplanes, changing along the $A$ axis should not affect its distance to $B$'s hyperplane. This can be achieved through projecting one direction onto another or optimization-based procedures. In practice, as shown in [20], the learned directions are often found as orthogonal yet entangled in the embedding vector space because projection can only remove linear correlations while nonlinear relationships between them could still exist. The second type of strategy [25] utilizes the information within the semantic mask and disentangles the features in different semantic regions. While this shows its effectiveness over more localized attributes, it fails to generate controls for global attributes like gender or age. Gradient-based channel filtering [8] selects channels based on the importance with respect to a target attribute. More concretely, by taking the gradient with respect to each attribute, Chen et al. [8] select the channels that have the maximal impact on the target attribute while filtering out the channels with the maximal impact on other attributes. However, this could fail if two attributes are strongly correlated and share almost the same set of channels for decisions.

What if the original GAN space is "entangled"? In this case, will the GAN model be able to generate images that never or hardly appear in the training data? For instance, images which are typically out of the distribution of the training data such as men with lipstick or women with beards. We hypothesize that the lack of such training data results in non-uniformly distributed density (hence entanglement) in the learned latent space, which leads to the bias of the identified directions. To support our intuition, we first show the empirical findings of the correlation between different attribute pairs in the original image distribution and how this affects the learned GAN space. Inspired by our empirical findings, we propose a novel framework called SC$^2$GAN to obtain disentangled controls. In particular, we project generated samples with GAN inversion methods back into the low-density regions in the learned latent space to achieve a more balanced latent space distribution, which can help decorrelate the pairs of attributes. We show that the interpretable directions re-learned by different methods under our framework would be corrected towards the correct cluster, as shown in the right panel of **Figure 1**.

Our contributions can be summarized as follows:

- We study the entanglement problem by showing the unbalancedness of certain attribute pairs in the original dataset and the resulting bias in the learned latent space.
- We propose a simple yet effective framework called SC$^2$GAN to disentangle features by correcting the biased latent space via projecting certain manipulated sample groups.
- We qualitatively and quantitatively show the effectiveness of our method, which is applicable as a post-processing procedure to many existing approaches that learn interpretable directions.

## 2. Related Work

We provide an overview of two different categories of approaches to control GAN outputs, as well as the line of work that embeds real images into GAN latent space.

### 2.1. Image Editing with Conditional GANs

By incorporating class label-related loss terms during training, conditional GANs obtain explicit controls over the generation process [17, 12, 19], which can generate images of classes specified by the user with a class label as input. Nevertheless, they lack controls over fine-grained attributes hence the entanglement issue can still occur. Recently, in the face image generation domain, new methods have been proposed to gain more fine-grained controls over multiple attributes [9, 21]. These approaches translate 3D face rendering controls, i.e., 3DMM [6] parameters, into the GAN framework, and are able to control the expressions, pose and illumination while preserving the identity. However, controls learnable by these methods are limited to existing 3D models' parametrization of facial attributes.

### 2.2. Interpolation in GAN Latent Space

Unlike conditional GANs, another line of work [4, 20, 11, 25] explores controls over output image semantics in GANs trained without labels. They have shown that such GAN latent space encodes rich semantic information with numerous meaningful directions, interpolation along which results in human-interpretable changes in the output semantics. InterFaceGAN [20] employs pre-trained image classifiers to cluster latent codes corresponding to different semantics and trains SVMs on those samples to learn the editing direction. Grad-Control [8] works similarly by training fully connected layers on a small amount of labelled latent codes, and taking the classifier gradient directions as the meaningful path in the latent space. GANSapce [11] works in an unsupervised way by performing PCA on features in the generator, and regressing linear directions in the latent space corresponding to the principal components, which correspond to human-interpretable changes in the image space. StyleSpace [25] learns more fine-grained controls by computing latent channels exclusively activated for semantic regions defined by pre-trained semantic segmentation networks.

Although various semantic directions have been discovered, during interpolation, entanglement in attributes, i.e.,

changing the target affects other causally independent attributes, often occurs. This phenomenon is known as spurious correlation [23, 3] and could be ascribed to the nature of the learned latent space, i.e., groups of visual attributes are not guaranteed to be uniformly distributed in the training data, hence the generator captures such spurious correlations and implicitly encodes them in its latent space. To address this issue, in the context of GAN latent space, Shen et al. [20] proposes to adjust the editing directions and minimize the change in the entangled attributes by orthogonalizing the target direction from the entangled attribute through projection, while [8] filters out salient latent channels for predicting the entangled attribute during interpolation. More generally, if the structure of the causal graph is known, Wang et al. [23] propose a post-processing method to provably identify and filter the feature subspace spanned by these spurious features by projecting the features to the so-called invariant feature subspace.

Although the aforementioned methods achieve partial success, disentanglement remains challenging when the correlation between attributes is significantly strong. After the adjustment, the resulting direction often only brings trivial changes in the target, or very few channels are left for the target attribute. [25] suffers less from the entanglement issue as it focuses on attributes belonging to localized semantic regions, but it lacks the ability to edit global attributes like age that require changes over the entire image. Different from approaches that directly adjust the biased directions, we propose to utilize such directions and debias the learned latent space by generating samples in low-density regions, and re-learn the editing directions based on the corrected latent distribution.

### 2.3. GAN inversion

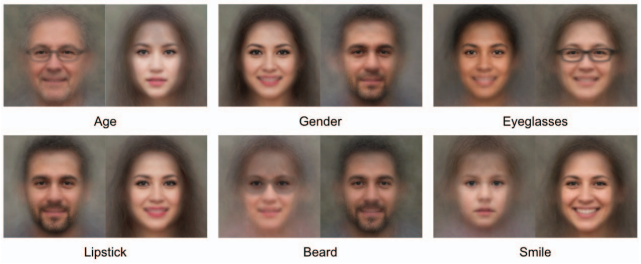
GAN inversion embeds real-world images into the GAN latent space, which can then be edited with latent space interpolation [1, 2, 15, 22, 24]. There are two main categories of GAN inversion: optimization-based methods [1, 15, 2], which sample from the original latent space and optimize the latent code to match the output with the real image target, or encoder-based methods [22, 24], which aim to invert the generation process and learn the reverse mapping from image space to the latent space, with the help from training on a large number of latent code-image pairs. One common challenge for GAN inversion is the tradeoff between distortion (i.e., resemblance to the target) and editability (i.e., how close the inverted code lies to the original latent distribution for the latent interpolation directions to be applicable), and different regularization methods have been proposed to handle such tradeoffs [28, 22]. In our work, we employ latent optimization to obtain samples with infrequent combinations of attributes, as we find it achieves little distortion and the results faithfully represent the minority distributions in the biased learned space.

## 3. Methodology

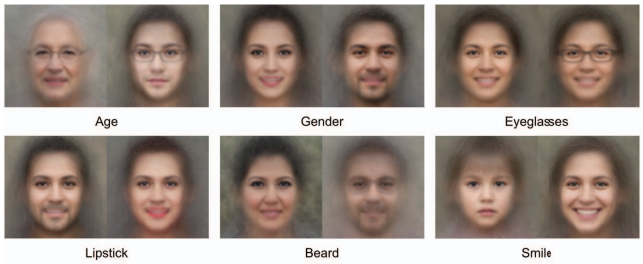
In this section, we describe the motivation and details of our method to self-correct StyleGAN $\mathbf{W}$ space and obtain disentangled image manipulations. We first show our empirical observations of entanglement in the learned latent distribution, followed by a more in-depth analysis that quantifies such phenomena. We then demonstrate our proposed novel framework that addressed the disentanglement by balancing the latent distribution. Specifically, we generate images that correspond to the low-density regions and rebalance the distribution by projecting those images back into the learned latent space.

### 3.1. StyleGAN Latent Space Entanglement

**StyleGAN Latent Space.** Our method to rebalance and disentangle the StyleGAN $\mathbf{W}$ space is motivated by observations of correlations between pairs of attributes in StyleGAN outputs. GANs map latent codes $z$ from a known distribution $Z \subseteq \mathbb{R}^d$ to an image space $X \subseteq \mathbb{R}^{H \times W \times 3}$ with the mapping function $g : Z \rightarrow X$. In StyleGAN [14], instead of directly feeding $z$ to the generative blocks, the out-

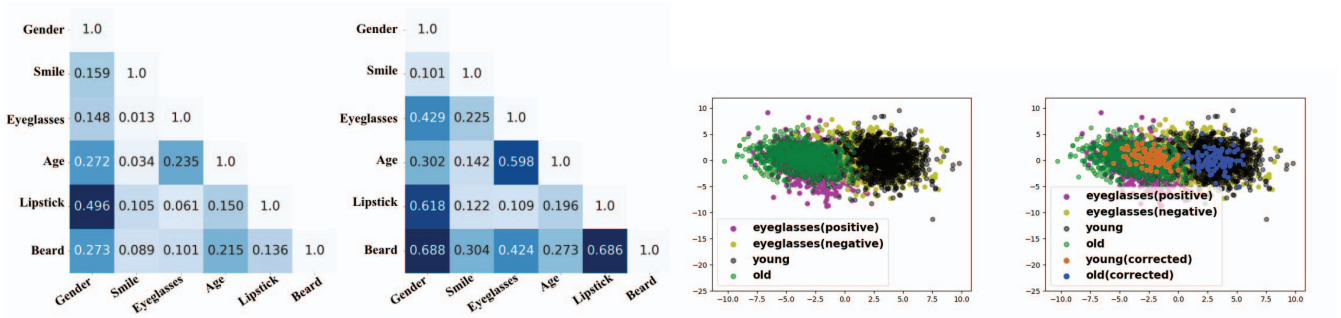



(a) Averaged faces for each attribute sign showing entanglement from the GAN-generated images.




(b) Averaged faces from the merged dataset containing equal amounts of original $\mathbf{W}$ samples and self-corrected samples. By projecting data onto low-density regions in the original clusters, e.g., edited images of old people not wearing eyeglasses or men with lipstick, the corrected distributions show less attribute correlations.

Figure 2: Averaged faces for each attribute sign sampled from the original $\mathbf{W}$ space and inverted from $\mathbf{W}+$ edits.




(a) Original FFHQ dataset. (b) StyleGAN2 generated images. (c) PCA of original. (d) PCA with corrected.

Figure 3: Illustration of the entanglement. (a) and (b) are the absolute values of tetrachoric correlations between each attribute pair in the original FFHQ dataset and StyleGAN-generated images. (c) is an example visualization of the latent space entanglement between eyeglasses and age, where the decision boundaries and interpolation directions for both attributes are likely to be similar. We show age clusters and eyeglasses clusters, fitted with PCAs trained on eyeglasses samples and projected with the top two components. In (d), we project the self-corrected samples (young with eyeglasses and old without eyeglasses) onto the same axes.

put is controlled by a function of $\mathbf{w} = M(z)$ where $M$ is a multilayer perceptron network with 8 layers. The vector $\mathbf{w}$ is a style vector lying in the $\mathbf{W} \subseteq \mathbb{R}^d$ space and each $\mathbf{w}$ vector is repeated 18 times to a $\mathbf{w}+$ vector and fed to generator layers at different resolutions to generate the final image with $G(\mathbf{w}+; \theta)$), which has been shown to enable powerful controls of features at different abstraction levels.

**Biases in learned W space.** Multiple works [14, 20, 11] have discovered that, unlike the original $Z$ distribution, the $\mathbf{W}$ space distribution is distorted as it captures the spurious correlations between attributes in the training data, resulting in low-density regions for the minority attribute groups. To visualize such effects, we randomly sampled 500k images from StyleGAN2-FFHQ [15] and employed pre-trained CelebA claasifiers [14] on our image bank to select the most confident 1k samples for a set of attributes. As shown in **Figure** 2a, we aggregate the 1k faces by computing the pixel space averages and show strong correlations among different attribute pairs, e.g., men with lipstick and women with beard are often underrepresented in the learned space, which directly relates to the entanglement problem many previous works $\mathbf{W}$ [20, 8, 11] suffer from, where editing one attribute affects the correlated attribute as well.

In order to further explain the findings above, we analyze the latent space entanglement from both the training data and the learned latent distribution perspectives. First, we analyze the original FFHQ training data for StyleGAN. With ffhq-features-dataset [18], **Figure** 3a measures the correlations between each pair of attributes, which exhibits non-trivial correlations between attributes like eyeglasses and age in alignment with our observations above. Next, we analyze the $\mathbf{W}$ space leveraging knowledge from pre-trained image classifiers [14]. In particular, with our 500k image bank and pseudo labels for each attribute of interest, we computed the same correlation matrix in **Figure** 3b. The big correlations between certain attribute pairs make the learning of disentangled editing directions challenging. For instance, for eyeglasses and age, since the high-density region for the old mostly contains old people wearing eyeglasses code samples, it's highly likely that when interpolating young latent code without eyeglasses following the old direction, eyeglasses will be added. Essentially, this corresponds to the overlaps between separation boundaries learned from data generated from the original $\mathbf{W}$ distribution, and we provide an intuitive explanation illustrated in **Figure** 3c. As discussed in [20], for such strong entanglement, orthogonalization of the editing direction through projection does not work well, as it also removes the target direction. Similarly, the salient channels proposed in [8] for both attributes also overlap significantly, making channel filtering prone to fail to disentangle the attributes.

**Disentangled edits with biased StyleGAN W directions.** Despite the biases in $\mathbf{W}$ space, previous works [11, 8] that learn editing directions based on $\mathbf{W}$ samples found a workaround to obtain disentangled edits. Instead of directly interpolating in $\mathbf{W}$, they apply the learned $\mathbf{W}$ directions to a subset of $\mathbf{W}+$ layers, which enables more localized controls. By limiting the changes to certain semantic regions, they successfully achieve disentanglement, e.g., only moving the mid-level $\mathbf{W}+$ layers to add lipstick and freezing the early layers that control global looks to avoid changing the gender.

However, $\mathbf{W}+$ interpolation has limited capacity as the changes are mostly limited to specific semantic regions. When editing attributes that involve global-wise deformation, limiting the changes to specific $\mathbf{W}+$ layers sometimes results in the desired target effect not being present, an example of which is shown in **Figure** 9. On the contrary,

$\mathbf{W}$ space modifies the image on a global level with greater ranges of changes available. Nevertheless, $\mathbf{W}+$ interpolation is still useful as it provides us access to StyleGAN-generated images with minority attribute groups. We hypothesize that, if these images can be reconstructed from latent codes in the $\mathbf{W}$ space, then such latent code cluster represents the low-density region needed for correcting the entangled $\mathbf{W}$ editing direction. Therefore, we ask the following question: Instead of manipulating the learned editing directions, if we could obtain low-density samples, e.g., old people without eyeglasses in the $\mathbf{W}$ space, and create a less biased (more balanced) training distribution for the editing direction, would the newly trained direction be more disentangled? An intuition behind our hypothesis is shown in **Figure** 3d where we aim to create a more balanced distribution for age clusters in terms of samples wearing eyeglasses.

### 3.2. Learning Disentangled $\mathbf{W}$ Directions

In order to debias the learned latent space distribution, efforts are needed to first identify the low-density regions and then acquire or generate the corresponding images. To automate this process and enable a large-scale correction, we introduce our method called SC$^2$GAN, which corrects the bias in the $\mathbf{W}$ distribution via self-corrected latent code samples. Given an entangled editing direction in $\mathbf{W}$, our propose to first interpolate $\mathbf{W}$ codes in $\mathbf{W}+$, which often shows more disentangled controls but does not generalize well for all attributes with the same hyper-parameters. Following such direction, we obtain edited images with localized changes corresponding to minority attribute groups, which will then be projected to the $\mathbf{W}$ space via GAN inversion. To enable disentangled editing, we re-train [20, 8] with this self-corrected latent distribution to learn the editing directions. We now describe the details of each step.

**Latent interpolation in $\mathbf{W}$ and $\mathbf{W}+$.** With an editing direction $f_a(\mathbf{w}) \subseteq \mathbb{R}^d$ learned, which could be a constant vector in the latent space, or a function of $\mathbf{w}$, interpolation in $\mathbf{W}$ space for 1 step with step size $s$ follows:

$$\mathbf{w}' = \mathbf{w} + f_a(\mathbf{w}) * s \tag{1}$$

To interpolate in $\mathbf{W}+$, the editing process can be denoted as

$$\mathbf{w}+^{'} = E(f_a, \mathbf{w}, \{i\}, n, s) \tag{2}$$

, where the $i$th $\mathbf{W}+$ layers are edited following **Equation 1** starting from $\mathbf{w}$ for $n$ steps with step size $s$. For example, $E(f_a, \mathbf{w}, \{0, 1, 2, 3\}, 3, 0.5)$ means moving only the first 4 $\mathbf{W}+$ layers for 3 steps with step size 0.5, while $E(f_a, \mathbf{w}, \{0..17\}, 3, 0.5)$ is equivalent to interpolating in $\mathbf{W}$ space. As observed in multiple previous works [25, 11, 8], compared to $\mathbf{W}$, the $\mathbf{W}+$ space from StyleGAN enables more localized controls, with $\mathbf{W}$ codes fed to layers at different resolutions controlling different abstraction levels, and the entanglement issue can be alleviated with spatial-wise editing in $\mathbf{W}+$.

**Obtaining Self-corrected Samples in $\mathbf{W}$.** To verify our hypothesis above, we employ a latent optimization process and project the disentangled $\mathbf{W}+$ interpolation results to the $\mathbf{W}$ space following:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \ell(G(\mathbf{w}; \theta), G(\mathbf{w}+^{'}; \theta)) \tag{3}$$

and we observe that the inverted $\mathbf{W}$ codes faithfully reconstruct the $\mathbf{W}+$ editing results and preserve the minority attribute groups well. In other words, these latent codes are self-corrected (disentangled) samples based on the original entangled editing directions, and they can be merged with the original $\mathbf{W}$ space samples to create a more balanced distribution for re-learning the editing directions. Assuming the original $\mathbf{W}$ space distribution has the entanglement issue between attribute $a_1$ and $a_2$, where the high-density regions mostly contain latent codes $\mathbf{W}$ with $(a_1-, a_2+)$ and $(a_1+, a_2-)$ semantics in the image space, hence changing the sign of $a_1$ by interpolating in $\mathbf{W}$ is likely to cause the opposite change in $a_2$. Through the process described above, we obtain $\mathbf{w}^*$ codes corresponding to images with $(a_1+, a_2+)$ and $(a_1-, a_2-)$ semantics in $\mathbf{W}$ space, hence the strength of $a_2$ in each $a_1$ cluster can be balanced by merging the original $\mathbf{w}$ codes with the self-corrected $\mathbf{w}^*$ codes. By retraining the editing direction for $a_1$ with the corrected distribution in $\mathbf{W}$, we decouple $a_1$ from the signs of $a_2$ and achieve disentangled and global controls.

## 4. Experiments

In this section, we apply our framework to existing supervised methods that learn editing directions based on $\mathbf{W}$ space latent code samples and obtain more disentangled directions. We first visualize such improvements by comparing the interpolation results before and after applying our framework, then quantify the amount of improvement for disentangling attribute pairs.

### 4.1. Experiment Setup

**Models.** We perform our experiments on the $\mathbf{W}$ space of StyleGAN2 [15] pretrained on FFHQ [14] with SVM-based [20] and gradient-based [8] editing directions. We sample 500k images and obtain pseudo labels for attributes gender, smile, eyeglasses, age, lipstick and beard with pretrained attribute classifiers [14].

**Learning the Original Directions.** Since our framework requires re-training of the learned editing directions, we first sample $\mathbf{W}$ latent codes corresponding to the images with the biggest/smallest logits from the classifier and follow [20] and [8] to learn the original $\mathbf{W}$ space editing directions.

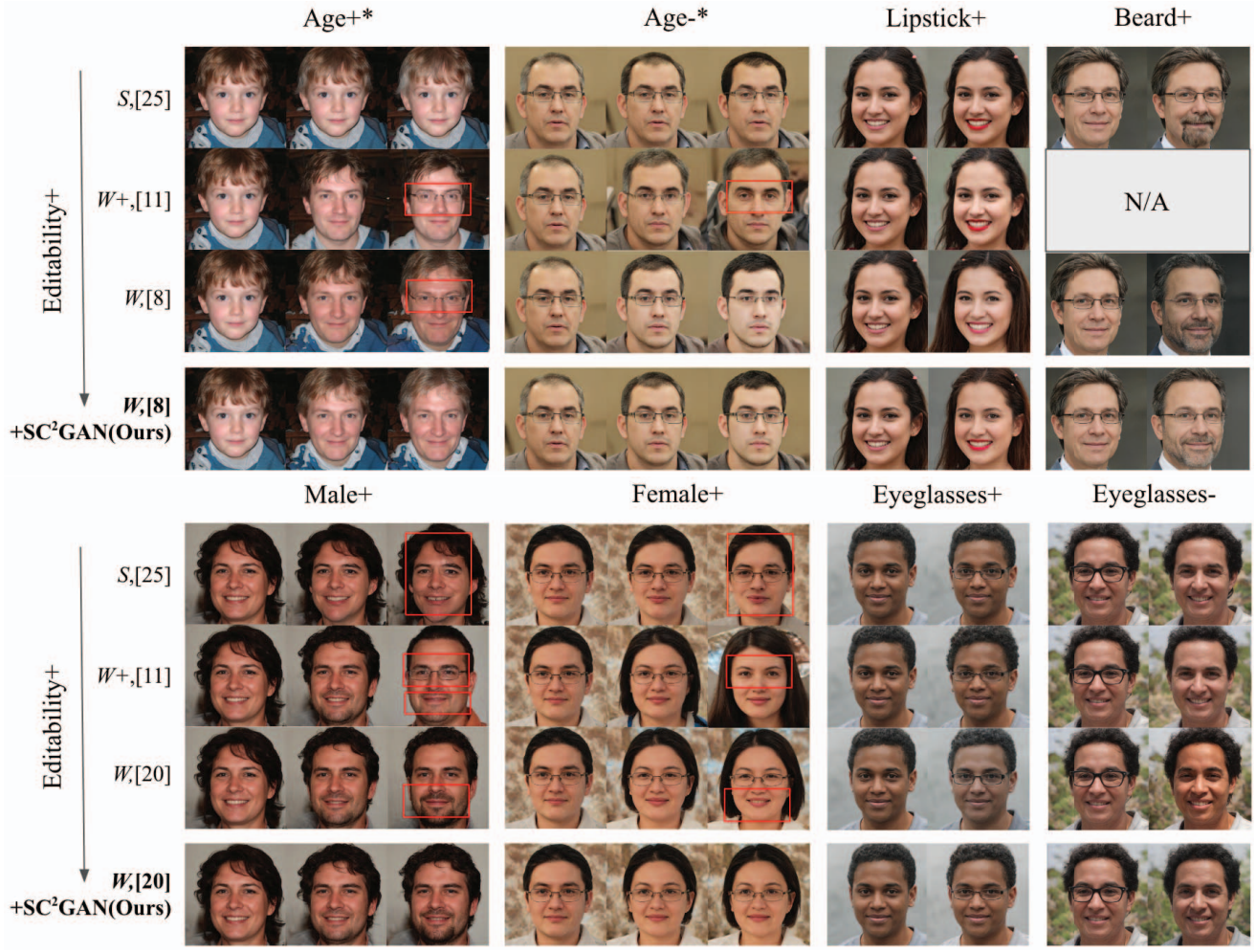



Figure 4: Disentangled editing results. N/A means unsupported direction. *Age for [25] is "white hair" and for the rest is "age". Within each group of images, left: source; middle, right: small and large interpolation direction. Our framework helps both methods [20, 8] obtain more disentangled $\mathbf{W}$ controls and achieve better results than $\mathbf{W}+$ [11] and $S$ controls [25] on global attributes like gender, and similar performance for localized ones like lipstick.

**Learning the Disentangled Directions.** With the original editing directions learned by each method, we apply **Equation 1, 2, 3** with the $\mathbf{W}+$ layer indices provided by [11, 8] to the set of training samples to obtain the self-corrected samples. We re-train the directions from scratch using the merged dataset containing both self-corrected samples and the original $\mathbf{W}$ codes. More implementation details can be found in the Appendix.

### 4.2. Disentangled Attribute Manipulation

We first present qualitative results for attribute manipulation for gender, age, eyeglasses, lipstick and beard in **Figure 4**, where we compare the original directions learned by Grad-Control [8] and InterFaceGAN [20] and editing directions after applying our framework to both methods. We also compare our results with methods to which our framework is not applicable. GANSpace [11] learns meaningful directions in $\mathbf{W}$ by applying PCA to generator features and requires manual examination for semantic meanings, while StyleSpace [25] finds locally activated semantic channels in $S$ space, which is $\mathbf{W}+$ layers with affine transformations applied. For both global attributes(age and gender) and local attributes(lipstick, eyeglasses and beard), our framework boosts disentanglement for both InterFaceGAN [20] and Grad-Control [8]. For instance, we achieve disentangled aging effects without eyeglasses added and decouple female direction from smile. GANSpace and StyleSpace suffer little from the entanglement issue, but the amount of change they make for global attributes is extremely limited, e.g., StyleSpace fails to synthesize more female effects, and

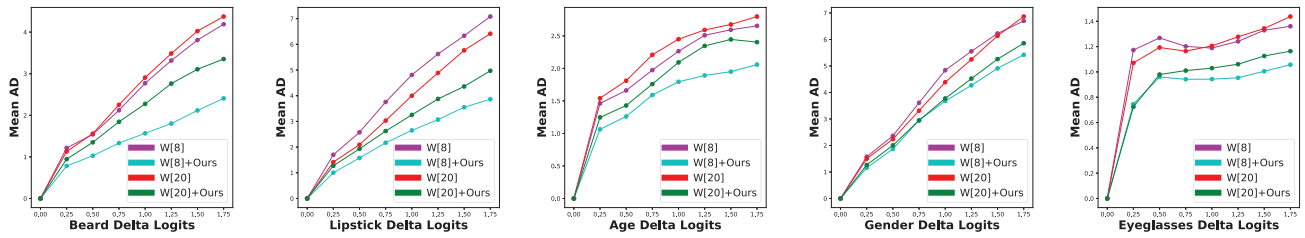



Figure 5: Attribute Dependency (AD), x-axis: normalized logit change in the target attribute, y-axis: mean of normalized logit changes in the others. Large y values mean strong entanglement. Our framework significantly reduces the amount of entanglement during interpolation in $\mathbf{W}$ space for both [20] and [8].

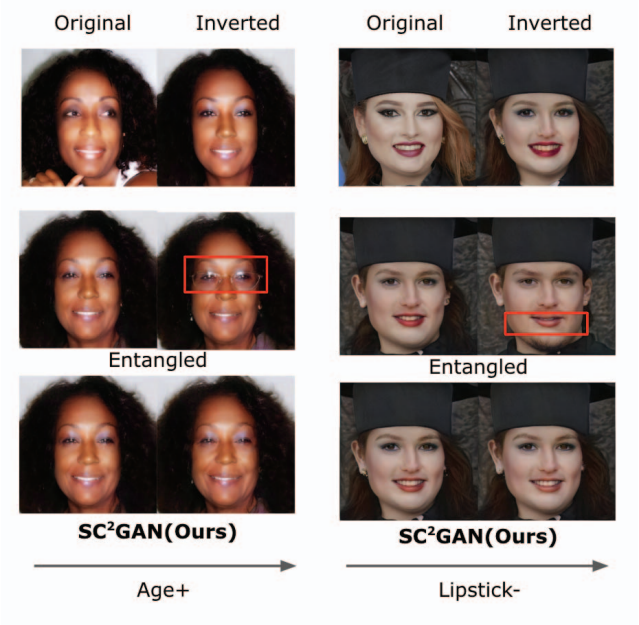



Figure 6: Disentangled controls with and without our framework applied to [20] for editing real images.

GANSpace lacks the ability to generate aging effects. In the meantime, for local attributes, with our framework applied, InterFaceGAN and Grad-Control achieve performance similar to GANSpace and StyleSpace, which operate in spaces of much higher dimensions.

### 4.3. Quantitative Results: Entanglement Analysis

We quantify the level of entanglement based on Attribute Dependency(AD) proposed by [25]. To compute the level of entanglement for one attribute $a$ with an editing method $f_a$, we first sample latent codes corresponding to images around the decision boundaries for the corresponding attribute classifier [14], and interpolate them with fixed step sizes $d$ for 9 steps. At each step $s$, we compute the absolute change in logits for the target $x = \Delta l_s^a$ and the sum of absolute logits changes in the rest of the attributes, divided by the population standard deviation of each attribute $y = \frac{1}{|A|-1}\Sigma_{i \in A \setminus a} \frac{\Delta l_s^i}{\sigma l^i}$, where $A$ stands for the set of all attributes. Finally, we group all points with respect to $(\frac{x}{\sigma l^a})$ into buckets of $(0, 0.25], (0.25, 0.5], \ldots, (1.75, 2]$, and plot the midpoint for each bucket as the final x-value, mean of $y$ values within each bucket as the final y-value. We append the full algorithm and more details in the Appendix. As shown in **Figure** 5, with our framework applied, the disentanglement in $\mathbf{W}$ editing direction improves significantly for [8, 20] on all attributes.

### 4.4. Real Image Manipulation

**Figure** 6 show the edited results on real images where entanglement exists for correlated features. Due to limited space, we demonstrate the age and lipstick edits following [20] with and without applying SC$^2$GAN. Our proposed approach achieves disentanglement while preserving the identity better.

## 5. Ablation Studies

**Number of Self-corrected Samples.** We qualitatively show how the number of self-corrected samples merged with the original $\mathbf{W}$ training data affects the overall editing directions learned by [8] in **Figure** 7, as their method can be trained only on a small dataset. With more self-corrected samples added, the original entanglement with eyeglasses is further minimized, with eyeglasses not appearing with similar aging effects present during interpolation.

**Directly Sampling Balanced Data.** An alternative approach that obtains the low-density area latent codes is to directly sample from such regions based on the pseudo labels of our image bank. However, as shown in **Figure** 8, although some entanglement can be alleviated with this approach, training with these samples could result in editing direction pointing to areas with lower image quality as the generator is not well-trained in those $\mathbf{W}$ regions. Furthermore, the amount of low-density data available for sam-

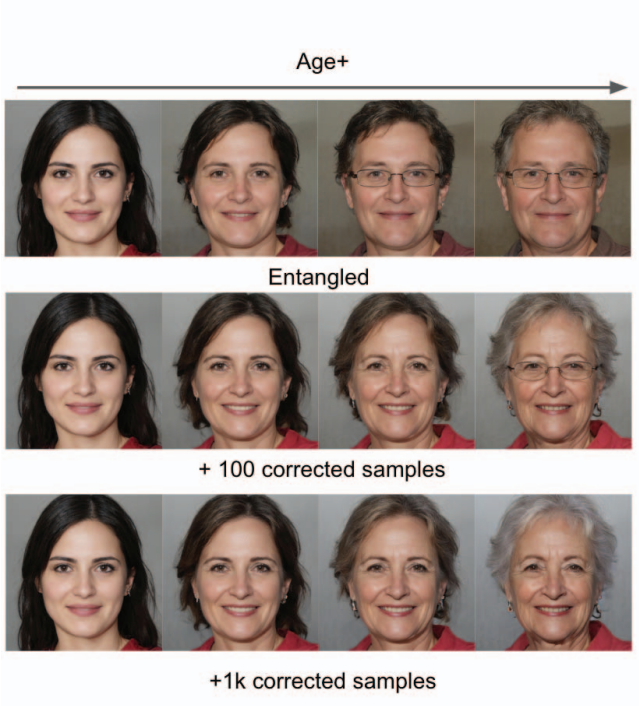



Figure 7: Comparison of different numbers of self-corrected samples added.

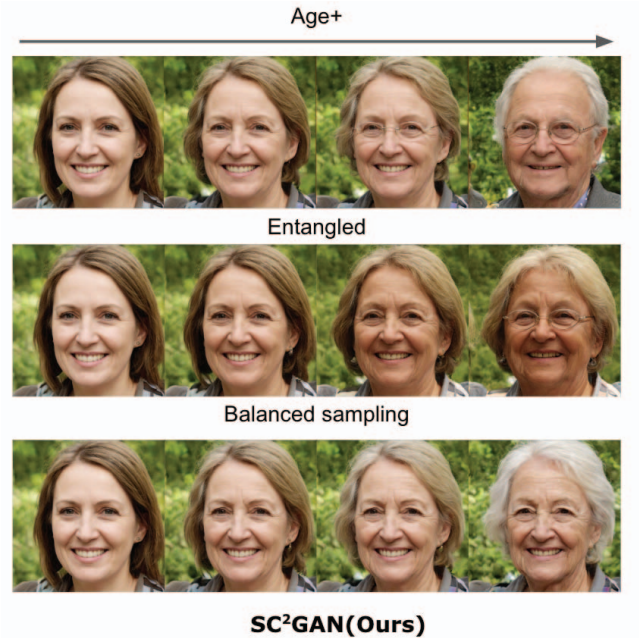



Figure 8: Comparison of balanced sampling from the original $\mathbf{W}$ space and our framework. Balanced sampling-based direction makes the image unnaturally dark.

pling is extremely limited, and as we take whatever is available given the scarcity of such data, these samples could lie close to the original separation boundary. Consequently, they may fail to provide a strong enough signal for the separation boundary to shift significantly.

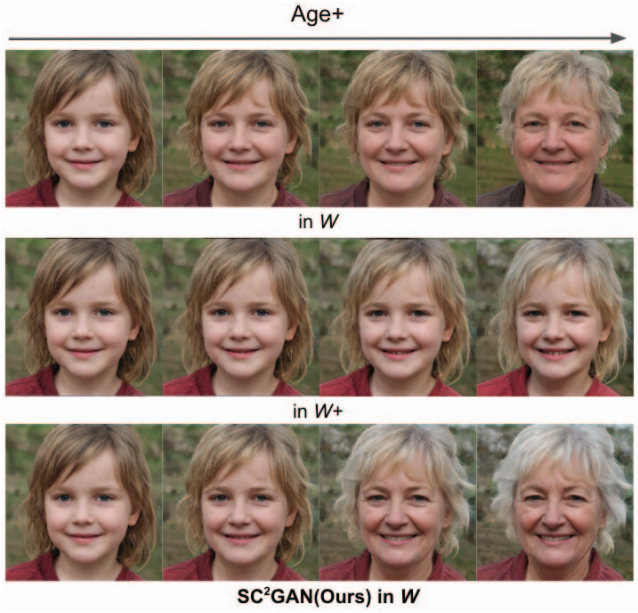



Figure 9: The difference between spatial-wise $\mathbf{W}+$ interpolation and $\mathbf{W}$ interpolation for increasing a child's age.

**Comparison with $\mathbf{W}+$ Space Editing.** We base our work on findings of [14, 11, 8] where $\mathbf{W}+$ space provides localized changes. Nevertheless, for attributes like aging, the editing involves great amounts of deformation of the original semantic regions, hence the localized $\mathbf{W}+$ space edits could fail to achieve the desired target effect, whereas interpolation in the $\mathbf{W}$ space is less prone to such failures as it modifies the image on a global level. We present an example in **Figure** 9. Both directions learned with our framework applied to [8] and the original direction with $\mathbf{W}+$ interpolation do not suffer from the entanglement with eyeglasses. Yet, the latter fails to create aging effects like saggy cheeks and ptosis of eyelids, with the changes mostly limited to the initial semantic regions.

## 6. Conclusion

We study the entanglement problem in the $\mathbf{W}$ space of StyleGAN2 and propose SC$^2$GAN, a simple yet effective method that generates self-corrected samples in low-density regions to obtain disentangled controls. With these self-corrected samples added to the original $\mathbf{W}$ distribution, we learn decoupled separation boundaries that enable disentangled editing. Overall, our framework shows strong capabilities to disentangle attributes with similar separation boundaries and salient channels in the original latent space, and works well in both local and global attribute manipulations.

## References


[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019. 3

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8296–8305, 2020. 3

[3] Gargi Balasubramaniam, Haoxiang Wang, and Han Zhao. Invariant feature subspace recovery for multi-class classification. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*. 1, 3

[4] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. 1, 2

[5] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. Seeing what a gan cannot generate. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4502–4511, 2019. 1

[6] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999. 2

[7] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. 1

[8] Zikun Chen, Ruowei Jiang, Brendan Duke, Han Zhao, and Parham Aarabi. Exploring gradient-based multi-directional controls in gans. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIII*, pages 104–119. Springer, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[9] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5154–5163, 2020. 2

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1

[11] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, 33:9841–9850, 2020. 1, 2, 4, 5, 6, 8

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 2

[13] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021. 1

[14] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1, 3, 4, 5, 7, 8

[15] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 1, 3, 4, 5

[16] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[17] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[18] mrmartin. ffhq-features-dataset. https://github.com/DCGM/ffhq-features-dataset, 2019. 4

[19] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, pages 2642–2651. PMLR, 2017. 2

[20] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 1, 2, 3, 4, 5, 6, 7

[21] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6142–6151, 2020. 2

[22] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 3

[23] Haoxiang Wang, Haozhe Si, Bo Li, and Han Zhao. Provable domain generalization via invariant-feature subspace recovery. In *International Conference on Machine Learning*, pages 23018–23033. PMLR, 2022. 1, 3

[24] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11379–11388, 2022. 3

[25] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12863–12872, 2021. 1, 2, 3, 5, 6, 7

[26] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE*

*international conference on computer vision*, pages 5907–5915, 2017. 1

[27] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1

[28] Peihao Zhu, Rameen Abdal, Yipeng Qin, John Femiani, and Peter Wonka. Improved stylegan embedding: Where are the good latents? *arXiv preprint arXiv:2012.09036*, 2020. 3