Applying StyleGAN image editing techniques to StyleGAN2-ADA

Pranav Bhagat
Department of Computational and Data Sciences
Indian Institue of Science (IISc)
pranavbhagat@iisc.ac.in

Abstract

This report presents an investigation into the potential of the latent space of StyleGAN2-ADA, a state-of-the-art data-efficient Generative Adversarial Network (GAN), for developing novel techniques for image editing. Our analysis revealed the existence of non-linear paths in the latent space that can be exploited for conditional attribute editing. Specifically, we found that manipulating certain attributes in the latent space can lead to significant changes in the generated image, such as altering the color of an object or adding or removing features. We developed a technique that allows for precise control over these attribute changes, enabling us to generate images with specific desired features.

1 Introduction

Generative Adversarial Networks (GANs) have revolutionized computer vision and image synthesis by learning to generate high-quality images that are nearly indistinguishable from real ones. However, GANs are typically dependent on large datasets, which can be expensive and time-consuming to obtain, limiting their applicability in fields such as medical and astronomy. To address this issue, an improved version of GANs called StyleGAN2-ADA (ADA-adaptive discriminator augmentation) has been proposed, which can be trained with limited data.

However, a major challenge with StyleGAN2-ADA is overfitting, where the discriminator learns the whole dataset, rendering its feedback to the generator meaningless, leading to divergent training. To overcome this, the adaptive discriminator augmentation scheme is used, which reliably stabilizes training and improves result quality when data is scarce. Nonetheless, this model still struggles with managing biased training datasets and generalizes poorly.

One of the most exciting applications of GANs is image editing, where GANs' latent space is used to manipulate images in various ways, such as changing the color of an object or adding or removing features. However, most existing image editing techniques rely on the assumption that the feature space is linear, which may not hold true for DE-GANs that use small datasets. Therefore, this report explores the non-linear latent space of StyleGAN2-ADA, to investigate its potential for image editing.

Specifically, we analyze the latent space of StyleGAN2-ADA and look for non-linear paths that can be used for conditional attribute editing. We also investigate how the ADA architecture affects the properties of the latent space, using techniques such as latent space interpolation along attribute directions. By doing so, we hope to demonstrate the unique capabilities of StyleGAN2-ADA for image editing and provide insights into GANs can be leveraged for this purpose.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

2 Previous related work

Previous research has explored the latent space of Generative Adversarial Networks (GANs) and how it can encode different semantic features. Some studies have observed the vector arithmetic property of latent vectors in a specific attribute direction in the latent space. Two notable works, INTERFACE GANs and GANSPACE, have motivated us to explore the latent space further.

Most modern research has assumed that the latent space of latent vectors is linear. INTERFACE GANs confirmed this assumption by showing that attributes are binary classified in space and that training support vector machines (SVMs) in that space produces a linear space. GANSPACE used unsupervised learning to find the direction of maximum variance using principal component analysis (PCA) and showed that maximum variance in facial features is present in the attribute directions.

In existing research on StyleGAN2, which is trained on a large dataset, linear space assumption makes interpolation easy along any attribute direction. However, we have used StyleGAN2-ADA, which is trained on a small dataset, and have found that the linear space assumption does not yield good results in this case. Therefore, in our work, we explore the non-linear nature of the latent space of StyleGAN2-ADA and investigate its potential for image editing. By doing so, we hope to provide insights into the unique properties of this GAN architecture and its latent space.

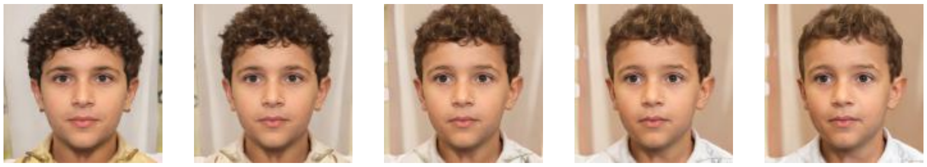


Figure 1: Space interpolation with stylegan2

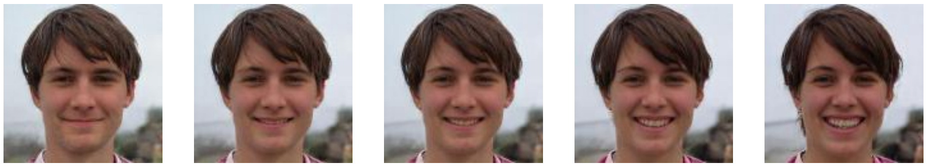


Figure 2: Linear interpolation in smile direction

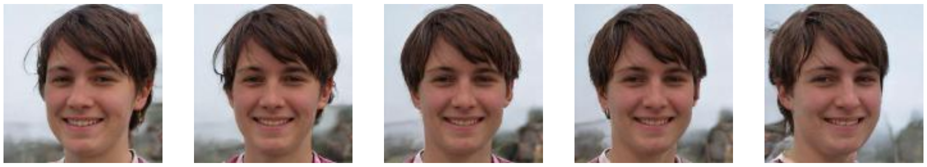


Figure 3: Linear interpolation in pose direction

These results obtained using linear interpolation in latent space of stylegan2. However, when we used latent space ADA for linear interpolation we got these results.

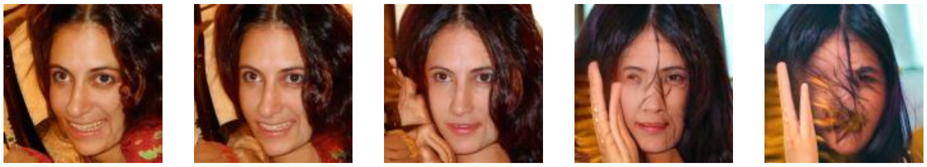
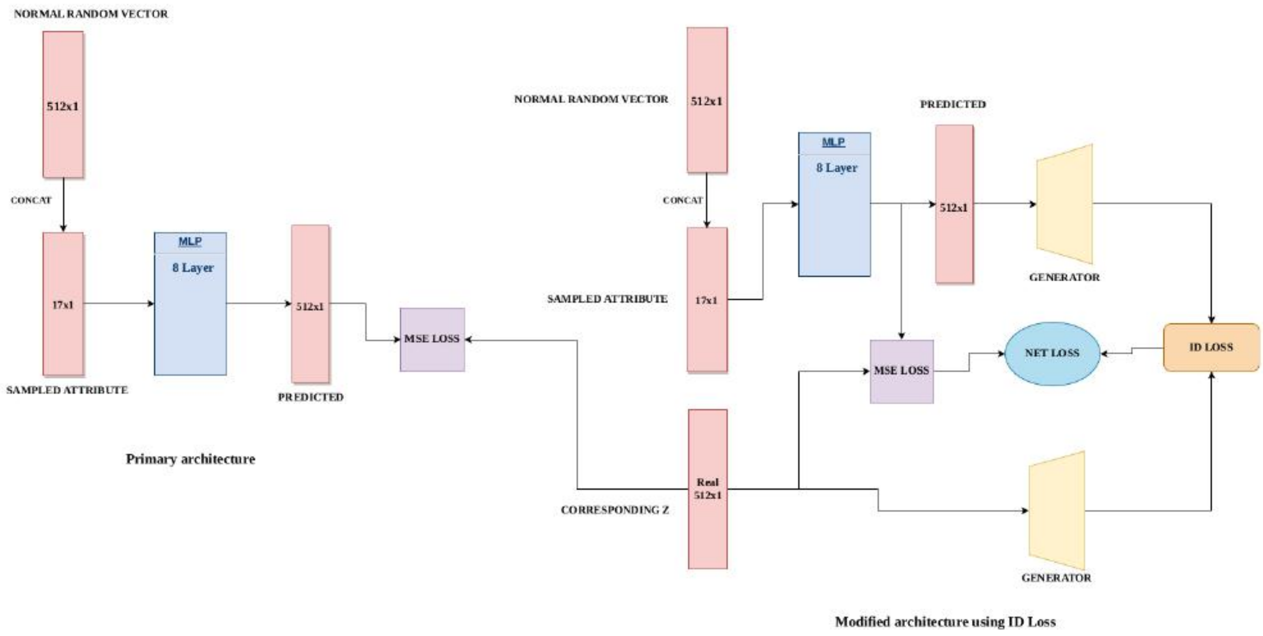


Figure 4: Linear interpolation in ADA space

As observed from above interpolation ADA performs poorly. Since ADA space is poor in semantics which results in poor generation with assumption of linear space.



Primary architecture

Modified architecture using ID Loss

Figure 5: Proposed architecture

3 Framework of our approach

In the context of conditional exploration of entangled latent spaces, we investigate the two sub-problems of attribute-conditioned sampling and attribute-controlled editing. We enable non-linear exploration of a GAN latent space and discover edit directions that are conditioned on the target images.

We support two tasks: first, attribute-conditioned sampling, wherein we want to sample high-quality realistic images with target attributes; and, second, attribute-controlled editing, wherein we want to edit given images such that the edited images have the target attributes while best preserving the identity of the source images.

3.1 Dataset

Preparing a dataset for image synthesis involves several steps. Firstly, we sample 10,000 images from the Gaussian space of StyleGAN2 and infer the corresponding disentangled *W* codes of the models. To avoid low-quality outlier faces, we truncate the vectors by a factor of 0.7, as suggested by StyleGAN. Using these *W* codes, we generate corresponding images *I(w)* via the StyleGAN2 generator, establishing a mapping between the *W* space and the image space *I*. To enable conditional control over image features, we use a face classifier network *A* to map the images *I* to the attribute domain *A*.

For training the non-linear network, we use triplets (*w*, *i*, *at*) where the attributes *at* do not depend on the variable *t*. The dataset is prepared using this workflow, and the attributes *at* we focus on in our work include gender, pitch, yaw, eyeglasses, age, facial hair, expression, and baldness/hair length. To prepare the *At* domain on the training dataset, Microsoft Face API (Microsoft, 2020), a state-of-the-art tool for face attribute classification was used. This API provides a diverse set of attributes, which are then used to train our network. For the lighting attribute, predictions used from the DPR model, which outputs a nine-dimensional vector per image measuring the first nine spherical harmonics of the lighting. This comprehensive set of attributes enables fine-grained control over image synthesis for faces.

3.2 Primary architecture

In our primary architecture, which involves training an MLP network with concatenated normal random vectors and attributes from our training set as inputs. The aim is to minimize loss against the corresponding Z vector. However, we found that the results obtained from this architecture were not as promising as we had hoped, as there was significant entanglement present between the attributes. To address this issue, we are considering implementing an ID Loss as an extension to our current network architecture.

3.2.1 Results using Primary Architecture

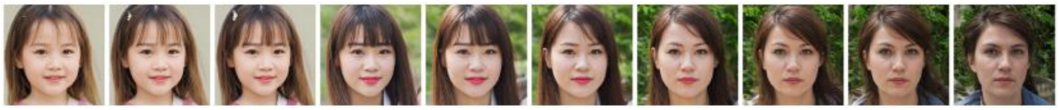


Figure 6: Variation along Age(gif)

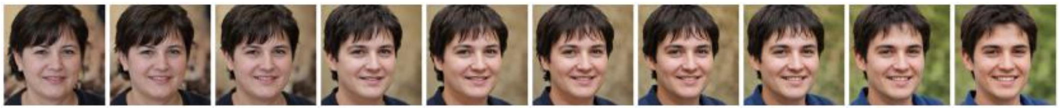


Figure 7: Variation along Gender(gif)

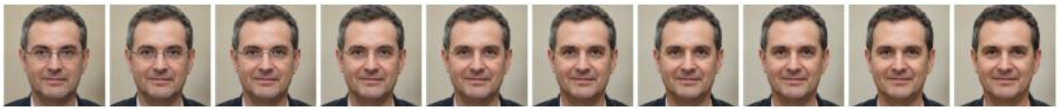


Figure 8: Variation along Glasses(gif)

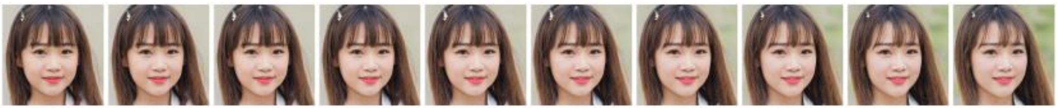


Figure 9: Variation along Pose(gif)

3.3 Architecture with ID Loss

After analyzing the results above, it is evident that the attributes are not completely disentangled. Multiple attributes are changing simultaneously, which indicates the need for an improved architecture for disentangled generation. To address this issue, we should consider using a more sophisticated model which consists of ID Loss, that can better disentangle the underlying factors of variation in the data.

The modified architecture includes an ID Loss, which is defined as the loss between an image generated with the real Z vector (from the training data) and an image generated from the predicted Z vector obtained from our MLP network conditioned on the corresponding attribute (Z). The total loss of our model is the sum of the ID Loss and the primary architecture loss. By incorporating the ID Loss, we aim to improve the quality of the generated images and ensure that they closely match the training data.

Loss = ID Loss + MSE

3.3.1 Results using Modified Architecture

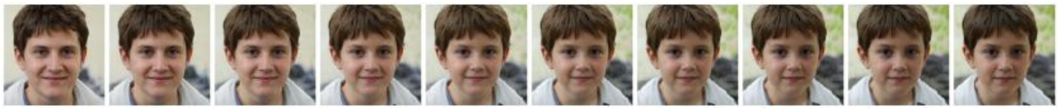


Figure 10: Variation along Age(gif)

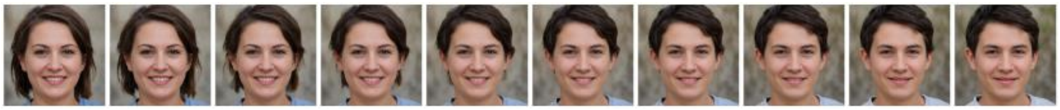


Figure 11: Variation along Gender(gif)

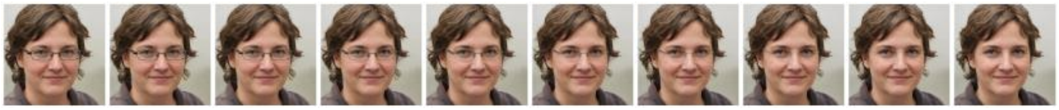


Figure 12: Variation along Glasses(gif)

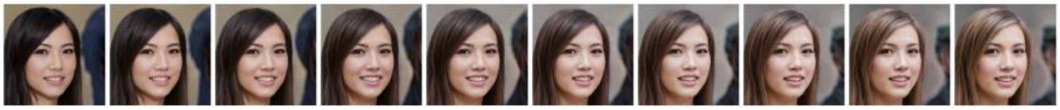


Figure 13: Variation along Hairs(gif)

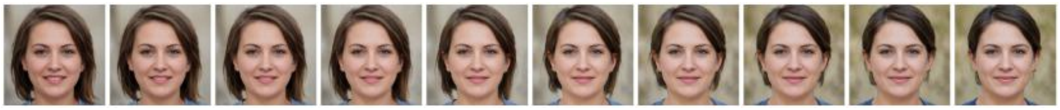


Figure 14: Variation along Smile + Hairs(gif)

4 Observation

In this section we will evaluate our experiments.

4.1 Manipulating Single Attribute

Our plots depict the results of manipulating five different attributes, indicating that our approach is effective in both positive and negative directions. Notably, when using the modified architecture, we observed that the latent code conditioned on attributes generated attribute-conditioned images. Despite a lack of training data for extreme facial poses, our GAN model was still able to generate realistic images of profile faces. These observations provide strong evidence that our GAN model is capable of learning interpretable semantics from the latent space, rather than producing images randomly.

Overall, these results demonstrate the effectiveness and versatility of our approach for attribute manipulation in facial images. By conditioning the latent space on specific attributes, we are able to generate images with controlled modifications to those attributes. Moreover, our model is able to extrapolate beyond the training data to generate images that are consistent with the learned semantics of the latent space. This makes our approach a promising tool for various applications in image editing and synthesis.

4.2 Conditional Manipulational

In this section, we study the disentanglement between different attributes and evaluate the conditional manipulation approach.

Facial attribute editing often involves multiple attributes that are semantically correlated, such as smile and expression or hair color and skin tone. To enable independent editing of such attributes, we propose a new method called conditional manipulation. The core idea behind this approach is to fix one attribute and then edit another attribute while holding the fixed attribute constant. This method allows for a more precise and targeted manipulation of specific facial features without altering other related attributes.

In addition, conditional manipulation allows for simultaneous editing of multiple attributes. Figure 14 illustrates an example of this, where both smile and hair length are edited simultaneously. By changing both attributes together, we can create more natural and realistic results that take into account the interplay between different features.

Overall, our proposed method of conditional manipulation provides a more controlled and flexible approach to facial attribute editing. By decorrelating different semantic features, we can achieve more precise and natural-looking results that better reflect the intended changes to the image.

5 User Study

To assess the visual quality and identity preservation of the edited images, we conducted a user study. In this evaluation, we presented participants with a set of attribute-edited images and asked them to identify the changed attributes. We also asked for feedback on attribute interpolation to further evaluate the effectiveness of our method.

The results of our study were generally positive for the modified architecture, demonstrating the effectiveness of our proposed approach for attribute manipulation. However, the results for our primary architecture were somewhat mixed, which was not unexpected given the limitations of the model.

Overall, our user study provided valuable insights into the effectiveness of our method and highlighted areas for further improvement. By incorporating user feedback and continually refining our approach, we can further enhance the visual quality and identity preservation of attribute-edited images.

6 Discussion

The edit paths in latent space of ours are non-linear. This is in contrast to StyleRig, InterfaceGAN, and GANSpace that use linear trajectories. Our evaluation indicates that this difference is not the only factor in explaining the edit quality of our method. It is also possible to set up our method approximating the nonlinear edit path by a linear one.

We identified three major limitations of our work. First, our work relies on the availability of attributes. These attributes might be difficult to obtain for new datasets and could require a manual labeling effort. Second, great results are only achievable with StyleGAN2-ADA trained on high-quality datasets, mainly FFHQ. It would be good to have different types of datasets of similar quality, e.g., buildings or indoor scenes, to better evaluate our method. The lack of availability of very-high-quality data is still a major limitation for evaluating GAN research.

7 Future scope

In future work, we plan to further improve the generalizability of our model by integrating the CLIP model encoder for text prompt encoding. This will enable us to condition the latent code of our model on text prompts, creating a text prompt image editing approach. We will apply the loss function in a similar manner to our previous approach, with the added benefit of being able to generate attribute-edited images based on text descriptions.

By leveraging the power of text prompts, we can create more complex and varied image edits, making our approach even more versatile and useful in a variety of settings. Overall, incorporating text prompts as a conditioning factor holds great promise for expanding the potential use cases and applications of our approach.

References

Abdal, Rameen, et al. "Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows." ACM Transactions on Graphics (ToG) 40.3 (2021): 1-21.

Härkönen, Erik, et al. "Ganspace: Discovering interpretable gan controls." Advances in Neural Information Processing Systems 33 (2020): 9841-9850.

Shen, Yujun, et al. "Interpreting the latent space of gans for semantic face editing." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

Patashnik, Or, et al. "Styleclip: Text-driven manipulation of stylegan imagery." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

Conde, Marcos V., and Kerem Turgutlu. "CLIP-Art: Contrastive pre-training for fine-grained art classification." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.