Disentanglement properties of Style-based Image Generation

Dhruv Agrawal
dagrawal@student.ethz.ch

Seyedmorteza Sadat
ssadat@student.ethz.ch

Sombit Sushant Dey
somdey@student.ethz.ch

January 14, 2022

Abstract

In this project, we apply the StyleSpace analysis on StyleGAN, StyleALAE and StyleFlow architectures to study the effect of the mapping network on the disentanglement properties observed in the latent spaces $\mathcal{W}$ and $\mathcal{W}^+$, and the Style Space $\mathcal{S}$. Furthermore, we study how these spaces compare to a disentangled space created using an invertible interpretation network from [4]. We compare the image generation quality and disentanglement between these models as well as the manipulation quality between Style Space $\mathcal{S}$ and StyleFlow.

1 Introduction

The fidelity of computer-generated images has improved significantly in recent years. In particular, the StyleGAN architecture [11] has become a popular method for generating photo-realistic face images. However, producing well-disentangled high quality images are still unsolved. To this end, Recent researchers have studied the disentanglement properties of different latent spaces of StyleGAN [14]. These methods focus on quantifying the level of disentanglement in StyleGAN latent spaces and finding disentangled channels to manipulate generated images. [14] focuses on the *Style Space* $\mathcal{S}$, spanned by the channel-wise style parameters. These work shows that this space is more disentangled compared to the intermediate $\mathcal{W}$ and $\mathcal{W}+$ spaces studied in prior works [1]. In this project, we compare the generative and disentanglement properties of different Style-based models trained on the FFHQ dataset [10] and investigate the quality of image manipulation in different latent spaces.

The remainder of this report is divided into 4 sections. Section 2 focuses on the Related Work and the descriptions of the models used in this project. Section 3 then discusses the disentanglement and generative power of the models using StyleSpace analysis. Section 4 presents the quantitative and qualitative results, and Section 5 mentions our conclusion and future work.

2 Related Work

StyleGAN Style-based image generation, first introduced in [10], has been the gold standard of high-quality image generation methods since the past few years. While traditional models produce an output image I from an input random noise z, StyleGAN first transforms z into an intermediate representation w with a mapping network $w = f(z)$ and generate the output image with a separate synthesis block $I = g(w)$ controlled with the style vector w through adaptive instance normalization (AdaIN) blocks. StyleGAN2 [11] further improves the quality of StyleGAN by replacing the AdaIN blocks with Modulated Convolution to remove the famous blob-shaped artifacts in StyleGAN images. Finally, StyleGAN3 further enhances the quality of StyleGAN2 outputs by making the generator invariant w.r.t input translation and rotation. In that sense, StyleGAN3 can offer similar output fidelity as StyleGAN2 while generating consistent interpolations (e.g. StyleGAN3 semantic details smoothly change according to their relative positions on the object surface when manipulating the input).

StyleFlow [2] formulates the problem of attribute-conditioned image generation and manipulation as an instance of conditional continuous normalizing flows in the StyleGan latent space. To solve the task of attribute conditioned sampling, the network samples z from a multi-dimensional normal distribution. A learned mapping function $\phi(z, a)$, where a denotes the attribute weights, is used to obtain the intermediate latent vector w. These weights are then decoded by StyleGan architecture to generate image samples that match the target attribute. $\mathcal{W}+$ space is used during image sampling and attribute manipulation which uses different w vectors for each synthesis layers (i.e. a 18x512 dimensional vector is used for image generation), dissimilar to the StyleGan2 where the same vector is repeated at all synthesis layers.

Adversarial Latent AutoEncoders (ALAE) [12] exploit the recent advances in GAN training to train an autoencoder for generating images. During training of the model, they simultaneously learn a Mapping Network F, a Generator G, an Encoder E and a Discriminator D. During inference, the mapping network F can be used along with Generator G to generate novel images. In addition, real images can also be encoded and manipulated using the encoder E and the Generator G. Using this method, the authors train an autoencoder *StyleALAE* with an MLP encoder and a StyleGAN based generator. This allows them to achieve generative quality close to StyleGAN as well as encoding and manipulating real images.

StyleSpace[14] studies the disentanglement properties of different latent spaces of StyleGAN2. They find that the latent space after the layer-wise affine transformation of $\mathcal{W}$ is more disentangled than the original $\mathcal{W}$ space itself. They call this latent space StyleSpace $\mathcal{S}$. They then find localized channels in $\mathcal{S}$ in order to manipulate images. These manipulations are more localized and result in more predictable manipulations.

We implement the StyleSpace analysis from [14] for different generative architectures (namely StyleGAN3 [9], StyleFlow [2] and StyleALAE [12]) to study the properties of StyleSpace $\mathcal{S}$ more generally. Furthermore, we compare the manipulation techniques described in Style Space against the StyleFlow architecture. Based on these experiments, we believe that we are able to show

1. StyleSpace $\mathcal{S}$ is more disentangled than Latent Space $\mathcal{W}$ for Style-based architectures.
2. Having layer-wise independant channels (e.g. an independent factor for each synthesis layer, or each style channel) is key to better disentanglement and manipulation properties.
3. Manipulating in StyleSpace $\mathcal{S}$ is more disentangled than StyleFlow but suffers from lower realism.
4. Concurrently training for Generative Quality and Disentanglement suffers from comparatively poor image quality. Hence, it might be a better idea to first optimize for image quality and then fine-tune the network for disentangled manipulations.

3 Methodology

We wish to study the correlation between Disentanglement and Generative Quality of different Style-based Generators. Therefore, we divide the methodology section into separate sections studying each of these properties for different models. Section 3.1 will focus on understanding the disentanglement properties of different model. Next, Section 3.2 focuses on the generative power of each network.

3.1 Disentanglement

While the current state of the art models can generate very photo-realistic images, we have little control over the attributes of the images generated. It is also desirable to be able to edit/manipulate these images using the generative models themselves. A disentangled latent space would allow the users to predictably and uniquely control different attributes of the generated images. Therefore, having disentangled latent spaces is studied very deeply [14, 1]. We study the disentanglement in 3 different ways:

1. Disentanglement Completeness Informativeness [3] metric
2. Perceptual Path Length
3. Qualitatively using Localized Channels

DCI The descriptions of each criteria for Disentanglement Completeness Informativeness metric can be found in [3]. We compare the DCI scores for each model in the $\mathcal{W}$ and StyleSpace $\mathcal{S}$ using 50000 samples. The results can be found in Table 2.

Perceptual Path Length PPL [5] measures the difference in the VGG16 embeddings of the consecutive images while interpolating between two random images generated by a model. We calculate the PPL value for StyleALAE and StyleGAN3 models using 50000 samples and use the previously reported values for the StyleGAN2. The results are shown in Table 3.

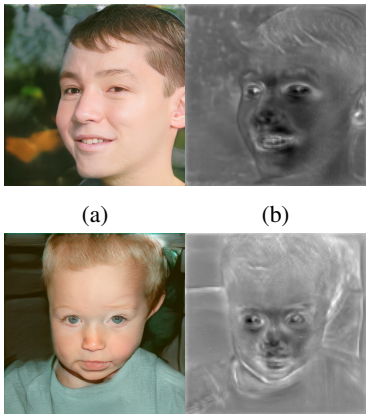
Localized Channels We use the methods described in [14] to identify localized channels in the Style Space $\mathcal{S}$ of different models. The results are discussed later in the report. Here, we describe some of the implementation details for identifying the localized channels. The generators used by the different models are similar. Each Synthesis block uses Convolution blocks and Affine Transformations. StyleALAE and StyleGAN 2 have 2 each while StyleGAN 3 has 1 each per Synthesis Block. However, the ToRGB blocks in StyleALAE do not use Affine transformations. They instead have a larger number of channels in the Synthesis Blocks. Secondly, the StyleFlow model learns a mapping for the $\mathcal{Z}$ space to a more disentangled $\mathcal{W}$ space. It then uses a pretrained StyleGAN2 Generator for generating images from $\mathcal{W}$ space. Therefore, we use pretrained Pytorch version of StyleGAN2 model from [8] repository. This also results in StyleGAN2 and StyleFlow having the same number of Style channels. The total number of Style channels differs between the models and is listed in Table 1.

3.2 Generative Quality

The most important metric for any generative model is the output image quality. In this section, we discuss the image generation quality of different models and the artifacts found in the StyleALAE using the FID score and visual comparisons.

Model	Number of Style Channels
StyleGAN 2	9088
StyleGAN 3	4950
StyleFlow	9088
StyleALAE	10176

Table 1: Different Number of Style Channels



(a) (b)

(c) (d)

Figure 1: Activation for StyleALAE

Fréchet Inception Distance FID [6, 13] score compares the statistics of generated samples to real samples. Lower FID scores indicate smaller difference between the generated and real distribution. The FID scores for different models on the FFHQ dataset is shown in Table 4.

Generative Quality of StyleALAE The generator of StyleALAE is insipred from the original StyleGAN generator. Therefore, it inherits some of its artifacts. Most importantly, StyleALAE follows the approaches of [7, 10] and uses progressive training. This might introduce high level signals to be captured in low resolution images and result in "phase" artifacts [11] in the high resolution images. In figure 1, we see that the activations at resolution 128×128 have high frequency details such as teeth shape and hair texture. Hence, we can conclude that the model certainly suffers from "phase" artifacts. One would solve this issues by removing AdaIN blocks and training the network without progressive growing.

4 Results

In this section, we discuss some of the results gathered from our experiments and their causes.

Model	Space	Disent.	Comple.	Inform.
StlyeGAN2	$\mathcal{W}$	0.6758	0.6208	0.9061
StlyeGAN2	$\mathcal{S}$	0.8018	0.8974	0.9524
StlyeGAN3	$\mathcal{W}$	0.4867	0.4270	0.9292
StlyeGAN3	$\mathcal{S}$	0.7694	0.8838	0.9845
StlyeFlow	$\mathcal{W}$	0.9730	0.8652	0.7858
StlyeFlow	$\mathcal{S}$	0.8890	0.8242	0.9565
StlyeALAE	$\mathcal{W}$	0.5868	0.5676	0.8574
StlyeALAE	$\mathcal{S}$	0.9270	0.7588	0.8590

Table 2: DCI Metric Results for all the Models $\times$ Spaces. Higher is better

Model	PPL (full)	PPL (end)
StyleGAN2	126.9	129.4
StyleGAN3	980.12	802.61
StyleALAE	181.83	114.94

Table 3: Perceptual Path Length Results for the $\mathcal{W}$ space. Lower is better

4.1 Metrics

The DCI, FID and PPL results for different models are listed in Tables 2, 4 and 3 respectively. Comparing the DCI scores, we clearly see that the Style Space $\mathcal{S}$ is more disentangled than intermediate space $\mathcal{W}$. The only model that breaks this rule is StyleFlow. We believe that this is because StyleFlow is specifically trained to disentangle $\mathcal{W}$. Applying Affine transformations to an already disentangled $\mathcal{W}$ space worsens the results.

Next, StyleALAE has better DCI and PPL scores than StyleGAN2 while having a significantly worse FID score. This indicates that StyleALAE is more disentangled but has worse image quality. This is also supported by the qualitative comparisons in Figures 2 and 4.

Lastly, the DCI and FID scores of StyleGAN 2 and 3 are quite similar. This is because StyleGAN3 solves translation and rotational artefacts from StyleGAN2 but these artefacts have a very minor effect on these metrics. However, adding latent translation and rotation representations hurts interpolation in latent space. This is also highlighted by a worse PPL score.

4.2 StyleSpace Manipulation

We manipulate samples from StyleALAE and StyleGAN3 using the Localized and the Attribute Dependent Channels found using StyleSpace analysis. The results can be seen in Figures 2 and 3 respectively. For the Lo-

Model	FID
StyleGAN2	2.84
StyleGAN3	2.79
StyleFlow	2.84*
StyleALAE	13.09

Table 4: FID Scores on FFHQ dataset. *StyleFlow uses the generator of StyleGAN 2 and hence has identical FID score. Lower is better.

(a) Hair

(b) Mouth

Figure 2: Localized Channels manipulation for StyleALAE

(a) Eyeglasses

(b) Oval Face

(c) Receding Hairline

(d) Age

Figure 3: Attribute Dependent Channels Manipulation for StyleALAE

calized Channels, we notice that the effects are very specific and the majority of the image remains unchanged. Both the subjects maintain the same expression and the background is largely unchanged as well. The effects of the manipulations are also very linear and predictable. However, the weakness of StyleALAE generation quality also gets highlighted as larger manipulations add unnatural artefacts in Figure 2 (b). Lastly, the in Figure 2 (a), the color of the background divided the hair of the subject.

In contrast, attribute dependency channels can effect larger areas of the image (Figure 3)but have more correlation between different attributes. For example, in Figure 3 both (a) Eyeglasses and (c) Receding Hairline effect the age of the subject. In contrast, manipulating (d) Age does effect the hair color or line.

We also manipulate the same attributes using StyleGAN3. The results are shown in Figures 4 and 5. We do not notice any changes in the background while manipulating the attributes. Furthermore, in Figure 4 (a), changing hair color is also able to change the small amount of hair on the shoulder. StyleALAE changes a smaller area of hair in contrast. However, in Figure 5 (a), the manipulation fails to add Eyeglasses to the subject.

In our testing, we found that Localized Channels are more effective in having the desired manipulations. Attribute Dependent channels are more likely to have imperceptible effect on the final image and have higher correlation.

Lastly, we can compare the manipulations with StyleFlow edits, which is specifically designed for attribute-based manipulations. The results are given in Figure6. We observe that Style Space manipulations are more localized but StyleFlow outputs are more realistic. This might happen because StyleFlow changes all style vectors of a given layer, i.e. multiple style channels at the same time, which results in more realistic outputs by preserving the relations between style-channels at the cost of more global changes. Hence, image quality seems to be at odd with perfect disentanglement

4.3 Image Manipulation using Invertible Interpretation Network (IIN)

As the $\mathcal{W}$ space of StyleGAN is highly entangled, we can use the IIN framework mentioned in [4] to transform this space into another disentangled space $\mathcal{W}'$. To this end, we train an IIN using a normalizing flow and Factor Loss to encourage disentanglement in the new space $\mathcal{W}'$. Then, we can first map the StyleGAN la-

(a) Hair

(b) Mouth

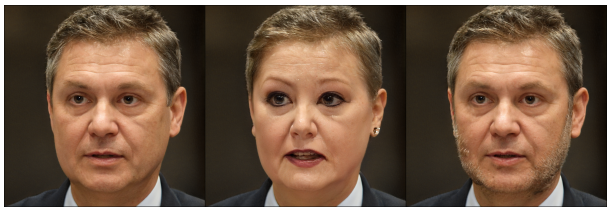
Figure 4: Localized Channels manipulation for StyleGAN 3

(a) Eyeglasses

(b) Male

Figure 5: Attribute Dependent Channels Manipulation for StyleGAN 3

(a) Editing attributes using StyleFlow



(b) Editing attributes using Style Space

Figure 6: Manipulations using StyleFlow and Style Space

(a) Global changes using modified w at all synthesis layers

(b) Localized changes using modified w only at the 5th synthesis layer

Figure 7: Image manipulation using IIN

tent vector $\mathcal{W}$ to $\mathcal{W}'$ and perform the semantic manipulations there. However, note that this naive approach results in global changes in the image because each vector w controls different styles at various resolutions. Hence, to get a localized edit, we change w in $\mathcal{W}'$ but when doing the inverse transformation, we change the input style of only one synthesis layer. The results are present in 7.

It is worth mentioning that this idea is similar to StyleFlow but instead of encoding the disentanglement in the network structure, it uses a specific loss function to encourage disentangled representation.

5 Future Work and Conclusion

In this work, we compared the disentanglement properties of style-based generative models and concluded that the Style Space $\mathcal{S}$ is generally more disentangled than the latent spaces $\mathcal{W}$ and $\mathcal{W}+$. We observed that using the Style Space, we can find localized channels that are responsible for semantic meanings in the generated output. However, manipulating images in the $\mathcal{S}$ space might reduce the realism of the output image. We also looked at two different methods, namely StyleFlow and IIN, to increase disentanglement of the latent space of StyleGAN. For future research, one can look into a combination of syle channels in each layer to increase the realism as well as effectiveness of Style Space manipulations. Furthermore, we can also train a different INN model to convert the $\mathcal{W}+$ space into a disentangled one to produce layer-wise control over the manipulations.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4431–4440, 2019.

[2] Rameen Abdal, Peihao Zhu, Niloy J. Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Trans. Graph.*, 40(3), May 2021.

[3] Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

[4] Patrick Esser, Robin Rombach, and Björn Ommer. A disentangling invertible interpretation network for explaining latent representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[5] Rani Horev. Explained: A style-based generator architecture for gans - generating and tuning realistic artificial faces. 2018.

[6] Neal Jean. Fréchet inception distance. 2018.

[7] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017.

[8] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Proc. NeurIPS*, 2020.

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

[10] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *CoRR*, abs/1812.04948, 2018.

[11] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020.

[12] Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto. Adversarial latent autoencoders. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. [to appear].

[13] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[14] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. *CoRR*, abs/2011.12799, 2020.