~:Summary for StyleGAN 2 (Karras et al 2020):~

Detailed Summary and Key Contributions for the paper : **Analyzing and Improving the Image Quality of StyleGAN (T. Karras et al. 2020)**

In broader terms, The baseline StyleGAN (SG1) was prone to bad quality images, small dimensions inputs and limited data was not an option. The same author came up with a new Generator architecture and training methods. Regularization of Generator architecture with this new methodology helps in generating high quality images. This regularization helps in better understanding of feature mappings to images. They improved baseline Style GAN with 6 additional modifications. Namely, Weight Demodulation, Lazy Regularization, Path Length Regularization, No Growing (G&D), Large Networks StyleGAN 2 (SG2)

The SG1 had various noises added during the final generation of the images. Moreover the author found that with the training data, SG1 is memorizing the places of the features for example, a face smile is placed in the face as if it is a sticker without noticing the face movement. Hence they proposed improvement in SG1 and further improving image quality. The paper only targets to improve the latent W. The noise which was added during the process was removed by changing the architecture with small tweaks. Improving the image quality is same as in ProGAN (T. Karras et al. 2017) and StyleGAN (T. Karras et al. 2019) but with an alternative architecture.

The intensive training from SG1 was relaxed in terms of calculating metrics continuously along with brute force approaches. The author proposed training in progressive minibatches (256,128,64,32 and so on). And also calculating FID and PPL after every 10 epochs, rather than calculating at each epoch. The metrics FID and PR are used to quantify the Classification criteria. However it was found that PPL Perceptual Path Length is a better metric which does not depend on textures rather than understands the shapes of the training samples and matches with generated images. They further mentioned that FID and PPL are the major metrics which are used in GANs. *Frechet inception distance (FID) measures differences in the density of two distributions in the high dimensional feature space of an InceptionV3 classifier. Perceptual path length (PPL) metric , originally introduced as a method for estimating the quality of latent space interpolations, correlates with consistency and stability of shapes.*

The small “artifacts” known as noise or water droplet-like features which were shown in the end results are balanced with two techniques. 1. Alternative Generator Architecture and 2. Alternative Normalization Technique. The author claims that showing such artifacts and historical features of the training process is quite puzzling that the discriminator should be able to identify that and detect it. However, the author studies about removal of such droplets by removing the normalization module from the SG2 architecture.

Modifications to Generator: Previously, in SG1 the weights and biases were updated in the styleblock of the architecture. It was observed that it creates an inverse relationship between weights & biases collectively and style features. Furthermore, for sake of improvement, they

moved the style block and the weight updates separately. In other words, essentially, it was expected that if we put weight updates and style features together, it will learn to reproduce better features and we can have features separated as we saw in SG1, however there was a problem observed very soon that when the weight updates for features are done together it created low quality outputs and discriminator was easily able to identify the fake images. The underlying purpose was not achieved. Hence, they moved the weight updates process outside the style feature block and operated directly on latent W data. In addition to that, they finally proposed that the weight and biases on input latent W can be removed without obstructing the resultant outputs.

Modifications to AdaIN: There was an observation of uncontrollable feature visibility when applied with different modulations and amplifications to specific details or features. One approach the author suggests is improving per image basis, which is actually very tedious. Author then proposed a new technique which was able to remove artifacts as well as preserving the freedom to amplify the details. The idea is to use a statistical approach for stabilizing the details. The actual mathematical derivation is given in the paper. But keeping in mind, the key idea is to use L2 normalization for the details. That is, whenever we are throwing an output from the generator we want to normalize that image such that its effect is balanced.

This statistical approach is quite a relaxation for the learning process because we don't have to memorize the positionings rather predict the behaviour of the final image distributions. The statistical analysis is a very neat approach used by modern researchers for not mugging up the whole image distribution rather than predicting the future positioning of details and features for faster convergence. One demerit is that, after every minibatch we have to observe the training process whether learning is in the right direction or not.

Modifications to Proressive Growing in SG2: Modifications were made to Instance Normalization and Progressive Growing in SG2 to control the features of the generated images. In practice, it was observed that the style mixing creates uncontrollable features in the final images which forced the generated images to memorize the feature positions. It is trying to generate strong correlation between details and other features of the image. (Recall that example of having a face smile at the same position even after face movements). Thus we are not maintaining or preserving the exact feature styles positions in the generated images, as well as not trying to come up with untrained feature styles.

In simpler words, the SG1 was trying to mug up the distribution of training data such that all the minute details like teeth orientation or nose positioning, are kind of learned on a frequency basis. The more the number of images with front facing smile the more frequently you will get the same smile orientation in generated images. This is not realistic as teeth orientation changes with face movement. Those kinds of features are called shift invariant features.

Reading Images from latent space: Earlier work says that it is quite difficult to project and find each image from its corresponding latent W image. Thus it will be much simpler to divide W layer wise and then search for our corresponding image. However, this will isolate those images

which cannot correspond to any latent W image but are equally useful. This method to project images is quite fast but bypasses nearly 20% of the generated images. Author proposes a way that we do not want to match any image in latent W. Instead we should look for latent W as well as unextended latent W as well which may have some images generated by the generator. In order to achieve this idea, they have proposed 2 changes:

1. Adding noise to latent W so as to explore latent space comprehensively.
2. Regularizing the stochastic noise inputs added.

Conclusions:

Author(s) have reviewed the earlier paper on SG1 (T. Karras et al 2019) and made some changes which was able to beat SOTA performance from the previous methods. They also proposed some better metrics to quantify the images. The training performance on the benchmark dataset was improved which were used previously.