EE782: Advanced Topics in Machine Learning Conditional StyleGAN for Audio Generative Modeling and Style Transfer

Anubhav Goel, 170040043
Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai, India
anubhav.goel@iitb.ac.in

Denil Mehta, 170100004
Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai, India
170100004@iitb.ac.in

Jian Vora, 170100026
Department of Electrical Engineering
Indian Institute of Technolgy Bombay
Mumbai, India
jianvora@iitb.ac.in

***Abstract*—This report documents the use of GANs for generative modeling of audio signals in conjunction with style transfer which is a relatively unexplored area. We study the concept of StyleGAN and its potential applications. We start with a model pretrained to FFHQ dataset to understand the working of StyleGAN and then proceed to benchmark its performance on other image datasets from different domains using FID scores. After obtaining good results across datasets, we extend the model to its conditional variant and apply it to an audio dataset by converting the audio clips to MEL spectrograms. The model generates audio clips with clearly distinguishable speech which leads us to conclude that the extension of StyleGAN to audio inputs is an effective technique to apply style transfer to this domain. Both qualitative and quantitative results are presented to validate our claims.**

***Index Terms*—style transfer, conditioning, MEL spectrograms, generative models**

I. Introduction

Generative Adversarial Networks have come a long way since their introduction in 2014. The main goal of these models is to try and model the underlying data distribution by training a generator which is adversarially trying to outperform a discriminator. GANs have been shown to produce high-quality images that are hard to distinguish even for humans, which has piqued interest in the community. A game-theoretic approach without modeling an explicit likelihood like VAEs, was a fresh idea bringing together concepts from various fields.

Given the success of GANs in modeling image distributions, we also evaluate how well GANs perform in the task of generating audio signals. StyleGAN is one of the recent works combining ideas from style transfer literature with GAN training and interpretable latent codes. We first analyze the working of StyleGAN on a variety of datasets (FFHQ, LSUN Cats, LSUN Cars). We extend the above model to implement style transfer to audio by mapping them to MEL spectrograms as an intermediate step and then tweaking the current architecture to include class conditioning the StyleGAN. We test our hypothesis on the Speech Commands dataset.

II. Related Work

In this section, we outline major works in three different domains which align the most with our goal. These include the neural style transfer literature, GAN literature and some recent work on generative modeling of audio.

A. Neural Style Transfer

Neural Style Transfer is the problem of taking a content image and a style image as input, and outputting an image that has the content of the former and the style of the latter. We refer to the first work in this space by Gatys et.al. [1] which minimized the sum of content and style loss by backpropagating on pixels. The style of the image is essentially the gram matrix of features extracted at various layers of pretrained models like VGG or Inception Net combined using weights to capture styles at various levels.

This was followed by a series of works trying to obtain faster versions of style transfer to work in real time mainly to be implemented as filters in various social media applications. Some of these include introduction of perceptual losses by Johnson et.al. [2], introducing adaptive instance normalization by Huang et.al. [3] and learning representations for artistic style by Dumoulin et.al. [4].

B. Generative Adversarial Models and StyleGAN

Generative Adversarial networks were first proposed by Goodfellow et.al. [5] as a model to learn functions to map isotropic gaussians to arbitrary distribution landscapes by using adversarial training. Numerous variants of Generative Adversarial Networks, such as InfoGANs [6] and Conditional GANs [7] have evolved over time introducing the ideas of interpretable latent variables and exerting control over various features of the images that are generated using these networks. One such variant which borrows ideas from the style transfer literature, known as StyleGAN [8], is analyzed in this paper. It has a series of fully-connected layers to map the noise latent vector $\boldsymbol{z}$ to a style vector $\boldsymbol{w}$. The generator takes inspiration from style transfer literature by including AdaIN (Adaptive Instance Normalization) blocks to modify the feature maps at each stage in accordance with the style vector. Thus training the architecture results in a map between random noise to a meaningful style which helps in interpolation in the latent manifold.

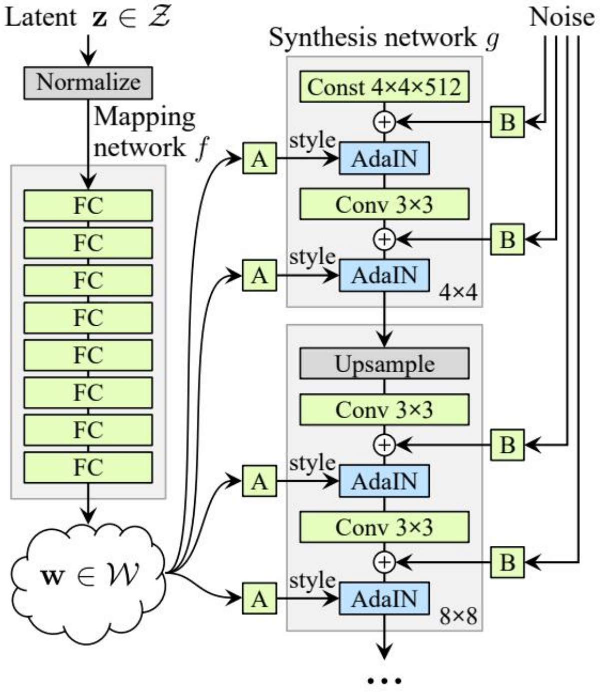

Fig. 1. StyleGAN Architecture

StyleGAN2 [9] searches for alternative designs that allow network design with great depth and good training stability. For ResNet [10], this is done with skip connection. So StyleGAN2 explores the skip connection design and other residual concepts similar to ResNet. It pinpoints the problem in the AdaIN operation that normalizes the mean and variance of each feature map separately, thereby potentially destroying any information found in the magnitudes of the features relative to each other and hence modifies these core blocks.

C. GANs for Audio Generative Modeling

Most work surrounding generative models has been limited to images and the success of GANs in this domain motivated people to apply this idea for audio generation. Most of the pioneering work in this space simply maps audio signals to a time-frequency plot called a spectrogram which is an image. WaveGAN by Donahue et. al. [11] was one of the initial works in this area by using a simple architecture similar to DC GAN for spectrograms. GANSynth [12] uses a Progressive GAN architecture to incrementally upsample with convolution from a single vector to the full sound. WaveNet [13] by Deepmind changes this paradigm by directly modelling the raw waveform of the audio signal, one sample at a time. As well as yielding more natural-sounding speech, using raw waveforms means that WaveNet can model any kind of audio, including music. The most recent work by Palkama et. al. [14] uses StyleGAN for style transfer in audio which is the core of this work.

III. Datasets

Generative Adversarial Networks generally require large image datasets to train. GANs fail at capturing multiple modes in the distribution, therefore, datasets should contain images of similar domains. GANs fail drastically for capturing the distribution of images in the wild such as in the case of ImageNet or Pascal dataset.

Our analysis involves the evaluation of performance on numerous datasets. We originally start with the **Flickr-Faces HQ (FFHQ)** dataset, which consists of 70,000 high quality images of 1024×1024 resolution. Compared to CELEBA-HQ dataset (the standard dataset of human faces), FFHQ is more diverse in terms of the background, ethnicity and age as well facial accessories such as sunglasses, hats, spectacles etc. Samples from this dataset are present in Fig 2.
Further, we train the model on a different dataset **LSUN Cats**, which has over 9000 images with the size 256×256 resolution. Samples from this dataset are present in Fig 3.
Finally, we train the model on the **LSUN Cars**, which consists of images with size 512×512. Samples from this dataset are present in Fig 4. For audio generation, we use the **Speech Commands** dataset which contains spoken digits uttered by multiple speakers in varying acoustic conditions. It also consists the class label for each audio clip which will be used by us in the conditional StyleGAN model. The dataset consists of 105,829 utterances of 35 short common words as a one-second or less WAVE format files. To reduce the complexity and computational constraints, we just use 18k samples each having fixed temporal length and a sampling frequency of 16kHz.

IV. Conditional StyleGAN for Audio Generation

As mentioned earlier, we implement the recent work by Palkama et. al. [14] on using StyleGAN for style transfer in audio as in Fig [6]. The basic pipeline is as follows: we first convert the wav audio clips into spectrograms on which the StyleGAN is trained. To compute mel-spectrograms from raw audio, we first generate the linear-frequency spectrogram with the short-time Fourier transform (STFT), then apply a mel-filterbank transformation to map the magnitudes to a mel scale, and finally convert the resulting mel-spectrogram to a decibel scale via a logarithm.

We also incorporate conditioning in this version of the StyleGAN, i.e., the style vector $\boldsymbol{w}$ is not just generated by the latent code $\boldsymbol{z}$ but also takes into account the class vector $\boldsymbol{c}$ which in our case is the class of the digit which is appended to the noise vector. Thus, we have the following,

$$\boldsymbol{w} = \boldsymbol{f}(\boldsymbol{z}, \boldsymbol{c})$$

For the discriminator, this class embedding $\boldsymbol{c}$ is passed to the hidden layers just like it is done for conditional GANs. Finally, audio clips can be generated from the spectrograms using standard python APIs like `librosa`.

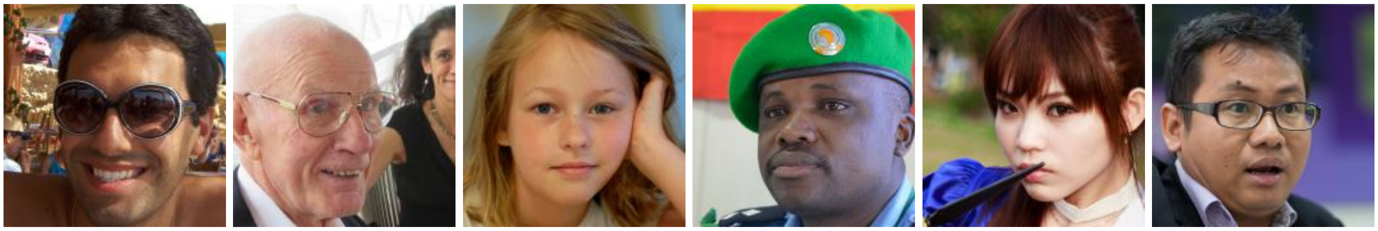

Fig. 2. Data samples from the FFHQ Dataset

Fig. 3. Data samples from the LSUN Cat Dataset

Fig. 4. Data samples from the LSUN Car Dataset

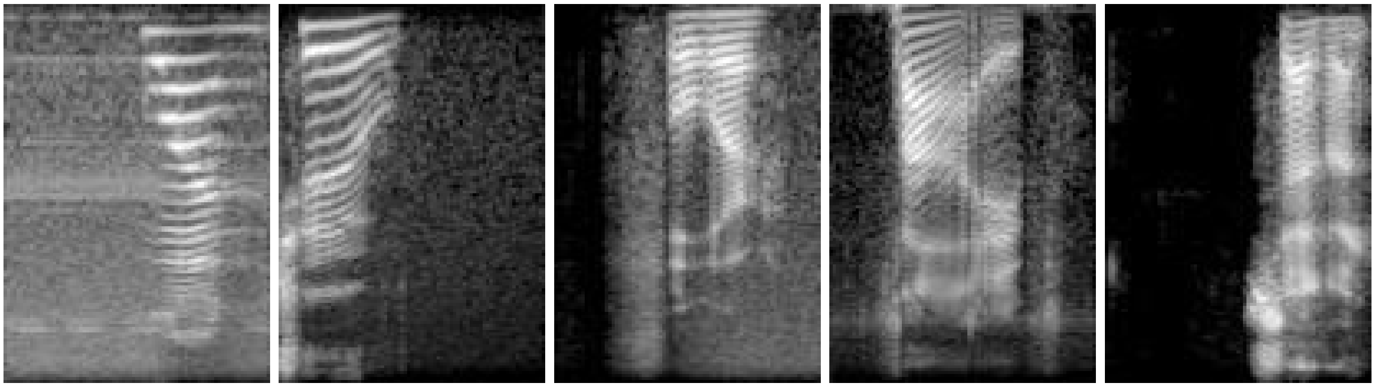

Fig. 5. Data samples from the Speech Commands dataset (converted to MEL spectrograms)

V. Analysis Pipeline

A. Testing and Benchmarking StyleGAN

As the first step, we analyze the original StyleGAN model across a variety of datasets which are commonly used for evaluating GANs. This was done to ensure a proper understanding of the model before trying the same for style transfer in audio. We inspect the samples generated both qualitatively and quantitavely in the form of FID scores. We also see transferring style by interpolating in the latent space of the trained model.

B. StyleGAN for Audio Generation

Once we have tested the basic StyleGAN model on standard datasets, we proceed to add the class conditioning feature in the model as decribed earlier and write the code for transformation of raw audio samples to MEL spectrograms. The latent dimensionality is 64 and we append a 10 dimensional class vector $\boldsymbol{c}$ as introduced in Section IV which implements a dense encoding for the digits 0-9. The resulting spectrograms of size 128×128 were fed into the conditional StyleGAN model. We use Adam optimizer to train the model with a learning rate $\alpha = 0.001$. The model was trained for 3 days on NVIDIA's 2080Ti GPU [see acknowledgments].

We then evaluate the model in a variety of conditions to test its correctness and robustness. We subjected our results to the following tests:

1) FID scores to ensure that the generated spectrograms' distribution is close to the original target distribution.

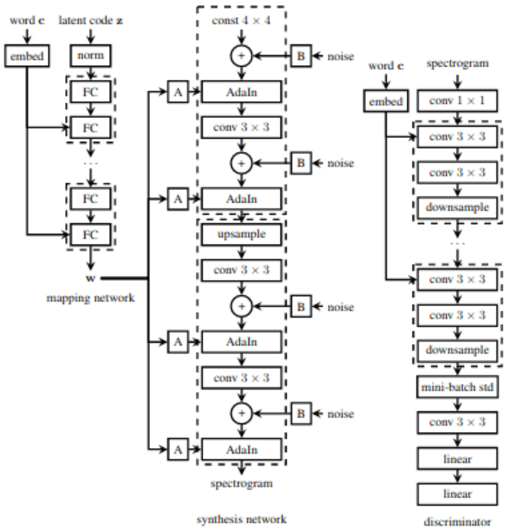

Fig. 6. Conditional StyleGAN Architecture

2) Qualitative evaluation of the audio generated by class conditioning, checking for noise and correctness.
3) Quantitative evaluation of the *coherence* of the generative model by measuring the classification accuracy of the generated spectrograms.

VI. Results

A. Testing and Benchmarking StyleGAN

First, we generate images from the FFHQ dataset. The generated images are shown in Fig 7. The grid shown in Fig 8 demonstrates the style transfer ability of the model. The style is taken from the first image of every row and is applied to the first image of every column to generate each image in the grid. The complete results are at the link https://drive.google.com/drive/folders/10su5FeHcc-_bAPhdnnQJ5B6ClD1RUK_o in the zip folder results_images.zip.

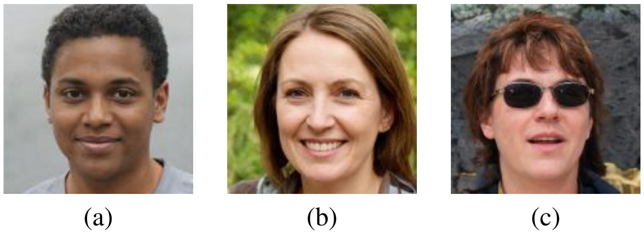
(a) (b) (c)

Fig. 7. Images generated after training on the LSUN Car Dataset

Dataset	FID Score	Baseline FID Score
FFHQ	4.42	4.42
LSUN Cats	7.12	6.93
LSUN Cars	7.37	6.69

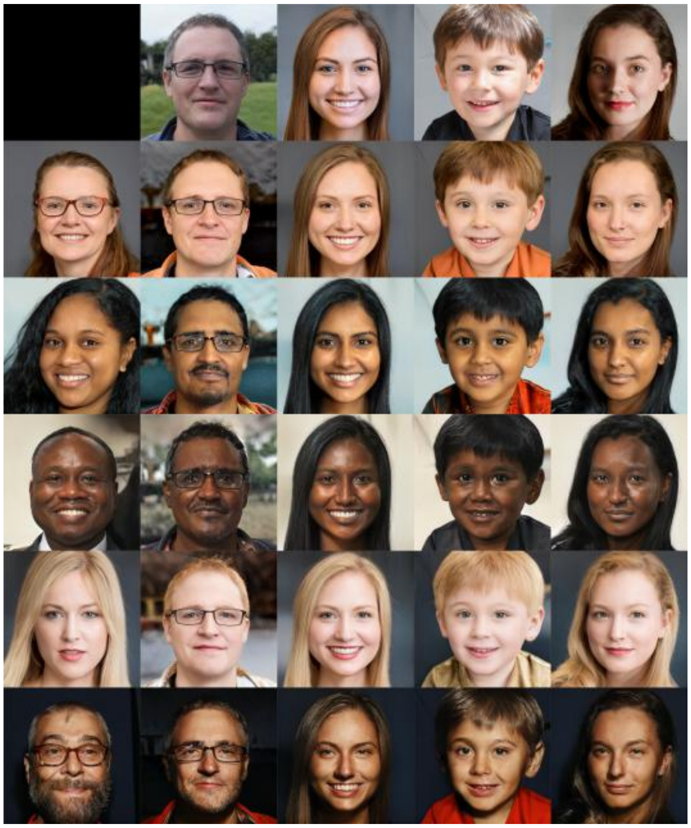

Fig. 8. Style mixing on FFHQ dataset

B. StyleGAN for Audio Generation

The spectrogram images generated by the conditonal StyleGAN have been presented below. Each image corresponds to a single digit from 0 to 9. The audio clips corresponding to each of these images have been included in the results folder in the submitted code. As the first metric, we evaluate the FID scores of the generated spectrograms which illustrates the closeness of two distributions given samples drawn from both of them. The FID scores came out to be **28.51** for $10,000$ samples. We built a classifier to classify the spectrograms into 10 classes. For the test dataset, we get an accuracy of 96.4%. To evaluate the correctness of the generator, we generate spectrograms and pass it thorough the classifier to get a multi-class accuracy of 88.4% which is very close to the baseline accuracy. Some generated samples of spectrograms are presented in Fig 9 and 10. The complete results are at the link https://drive.google.com/drive/folders/10su5FeHcc-_bAPhdnnQJ5B6ClD1RUK_o in the zip folder results_mels_jpg.zip and results_wavs.zip.

VII. Discussion

It can be observed that StyleGAN produces good results on the three image datasets which were used. The pretrained model works extremely well on the FFHQ dataset with almost all the generated images being indistinguishable from real human faces. However, as can be seen from the given

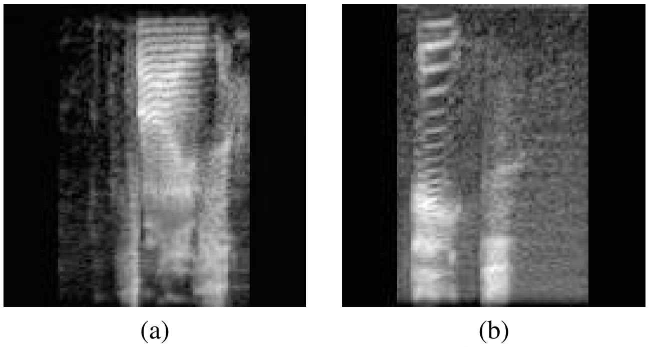
(a) (b)

Fig. 9. Spectrogram images generated for digits 5 and 8 respectively

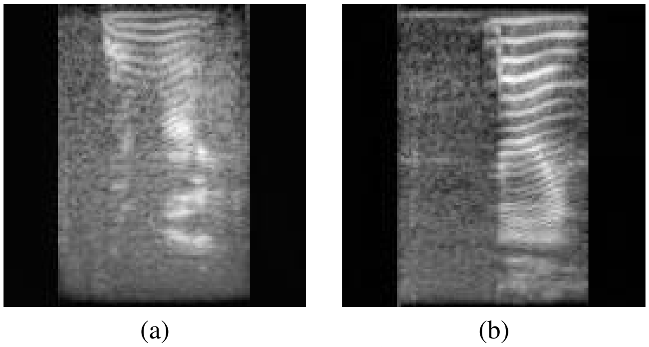
(a) (b)

Fig. 10. Spectrogram images generated for digits 1 and 9 respectively

generated image of the cat (Fig 11), we can see that sometimes StyleGAN does not work well on a new dataset. Another similar failure is shown on the cars dataset in Fig 11. However, these failures are extremely rare and generally the model performs well on new datasets as well.

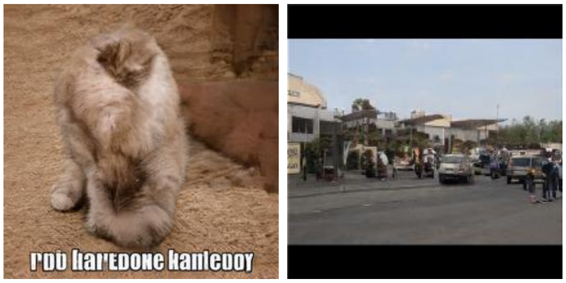

Fig. 11. GAN failing on LSUN Cat dataset and LSUN Car dataset

This motivated us to apply the concept of StyleGAN to an audio dataset, using the conditional StyleGAN. The results obtained in this case turned out to be accurate. The noise levels in the generated audio clips are low and the numbers can be distinguished clearly.

VIII. Future Work

Conditional StyleGAN works extremely well on audio generation. However, style mixing using the above architecture to apply a style from a source object to a target object is still an unexplored domain. This could be possibly serve as the first step in extending the use of this model in the audio domain. Another extension of the above architecture can be the use of more complex classes than digits. A convolutional architecture can be separately implemented to extract class information from images and these can serve as input to the conditional StyleGAN model. This will train an end-to-end model which takes as input an image and produces an audio clip describing the image in one or two words.

As a more complex problem, if the image and audio domains are unpaired and we want to perform style transfer, then we can take inspiration from the CycleGAN architecture to explore models along the lines of Cycle-Conditional-StyleGAN architecture.

References

[1] Leon A. Gatys, Alexander S. Ecker, Matthias Bethge, *A Neural Algorithm of Artistic Style*, arXiv:1508.06576 [cs.CV]

[2] Justin Johnson, Alexandre Alahi, Li Fei-Fei, *Perceptual Losses for Real-Time Style Transfer and Super-Resolution*, arXiv:1603.08155 [cs.CV]

[3] Xun Huang, Serge Belongie, *Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization*, arXiv:1703.06868 [cs.CV]

[4] Vincent Dumoulin, Jonathon Shlens, Manjunath Kudlur,*A Learned Representation for Artistic Style*

[5] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, *Generative Adversarial Nets*, arXiv:1406.2661 [stat.ML]

[6] Xi Chen, Yan Duan, Rein Houthooft, John Schulman†, Ilya Sutskever, Pieter Abbeel, *InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets*, arXiv:1606.03657 [cs.LG]

[7] Mehdi Mirza, Simon Osindero, *Conditional Generative Adversarial Nets*, arXiv:1411.1784 [cs.LG]

[8] Tero Karras, Samuli Laine, Timo Aila, *A Style-Based Generator Architecture for Generative Adversarial Networks*, arXiv:1812.04948 [cs.NE]

[9] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, Timo Aila, *Analyzing and Improving the Image Quality of StyleGAN*, arXiv:1912.04958 [cs.CV]

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, *Deep Residual Learning for Image Recognition*, arXiv:1512.03385 [cs.CV]

[11] Chris Donahue, Julian McAuley, Miller Puckette, *Adversarial Audio Synthesis*, arXiv:1802.04208 [cs.SD]

[12] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, Adam Roberts, *GANSynth: Adversarial Neural Audio Synthesis*, arXiv:1902.08710 [cs.SD]

[13] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, *WaveNet: A Generative Model for Raw Audio*, arXiv:1609.03499 [cs.SD]

[14] Kasperi Palkama, Lauri Juvela, Alexander Ilin, *Conditional Spoken Digit Generation with StyleGAN*, arXiv:2004.13764 [eess.AS]

[15] All starter code that was used has been referenced in the README file in the code repository