ATML Report

Alain Pulfer
alain.pulfer@irb.usi.ch

Brian Pulfer
brian.pulfer@usi.ch

Abstract

Recently, Viazoveskyi et al. suggested that although paired datasets were impossible to collect in reasonable times for tasks such as aging or gender swap, the usage of synthetically generated images could mitigate this problem. In their work, they genereate source images by employing a pre-trained StyleGAN2 on the FFHQ dataset. From each source image, they generate multiple candidate targets by repeated manipulations of the source latent vector. Candidate targets are filtered according to a custom network that retains only realistic manipulations and, finally, the resulting paired dataset is fed into the end to end Pix2PixHD network. Pix2PixHD is a conditional adversarial network used for image-to-image translation tasks that, unlike most conditional GANs, can generate photo-realistic images that have a high resolution, up to 2048 × 1024 pixels. Remarkably, Pix2PixHD networks trained on selected StyleGAN2 generated images produced better results in the gender swap, style-mixing and rejuvenation tasks compared to StyleGAN2 latent space manipulation and embedding. In this work, we are re-implementing the work of Viazoveskyi et al. with a focus on style mixing. Finally we are also evaluating possible extensions to this approach by generating pairs of synthetically generated faces and drawn portraits.

1 Introduction

One major drawback of classical GANs is represented by the entanglement of their latent space. In fact, while learning to generate realistic images from randomly generated vectors, these architectures would not necessarily learn a continuous mapping of the latent space to the image space. A small perturbation in the latent space would often result in a drastic change of the image content. StyleGAN architecture is designed to overcome this limitation by reducing the entanglement of the latent space. Mapping the original latent vector into intermediate latent vectors has the effect of creating a latent space W which is more disentangled and better represents the original distribution. This allowed to reach an unprecedented level of fine tuning of the image content by latent space manipulation, producing outstanding results in tasks such as ageing and rejuvenation, gender swap and style-mixing. StyleGAN2 further enhanced the first version by modifying the generator normalization, revisiting progressive growing and regularizing the generator in order to obtain good conditioning in the mapping from latent codes to images.

However, employing StyleGAN architectures for image editing has limitations. The first one is related to the computational cost associated with each image editing, as the real image needs first to be embedded in the latent space of the trained network. The second limitation is that image editing by latent space manipulation can be a trial and error process before reaching a visually pleasant and realistic result. The usage of an end-to-end networks can bypass these step by directly mapping the source image to a target image. However, the lack of ground-truth images, and therefore a validation, for morphing tasks such as gender swap or ageing is a major limitation for the training and end-to-end architecture.

Recently, Viazoveskyi et al. suggested that although paired datasets were impossible to collect in reasonable times, the usage of synthetically generated images could mitigate this problem. In their work, they genereate source images by employing a pre-trained StyleGAN2 on the FFHQ dataset. From each source image, they generate multiple candidate targets by repeated manipulations of the source latent vector. Successively, candidate targets are filtered according to a custom network that retains only realistic manipulations. Finally, the resulting paired dataset is fed into the end to end Pix2PixHD network. Pix2PixHD is a conditional adversarial network used for image-to-image translation. Unlike most conditional GANs, Pix2PixHD can generate photo-realistic images that have with high resolution, up to 2048 × 1024 pixels. Remarkably, Pix2PixHD networks trained on selected StyleGAN2 images produced better results in all the evaluated morphing tasks compared to StyleGAN2 latent space manipulation and embedding with perceptual loss.

In this work, we are re-implementing the work with a focus on style mixing. Finally we are also evaluating possible extensions to this approach by generating pairs of synthetically generated faces and drawn portraits.

Advanced Topics in Machine Learning, Prof. Cesare Alippi, SA 2020-2021

2 Related works

Generative adversarial network has been extensively used for the task of image to image translation. Due to their computational cost, large GANs model are often compressed to simpler student generative models (Aguinaldo, 2019. Chen et al., 2020)[1][2].

Others approaches make use of synthetically generated dataset for training classification models. In their work, Besnier et. al. [3] study the possibility of training a classifier only on generated data, and propose new techniques to obtain better classifiers exploiting the generative network.

Heljakka et. al. recently introduced the notion of automodulators [4], i.e. a new category of autoencoders that allow for style-mixing and other applications traditionally unreachable with traditional autoencoders. With the use of novel training techniques, they were able to produce high quality images outperforming traditional autoencoders in terms of visual image quality.

Multi-style transfer (MST) is the act of combining one content image with two style images. In their work, Huang et. al. [5] propose the first MST framework to automatically incorporate multiple styles into one result based on regional semantics.

Our work aims at reproducing the study that was conducted in [6]. In the paper, the authors use the vanilla Pix2PixHD network [7], an improved version of the work in [8] that can generate high-resolution photo-realistic images, to train an image-to-image model in applying aging, rejuvenation and style mixing. The training data is collected by exploiting the StyleGAN2 network generative capabilities [9] [10]. Our work focuses on the style mixing application.

3 Methods

Viazoveskyi et al. work is mainly based on the usage of a pre-trained StyleGAN2 architecture and different latent space manipulations [6]. In fact, the addition of a direction vector to a latent vector has the effect of shifting the corresponding generated image towards the desired feature, while subtracting it has the opposite effect.

In their study, Viazoveskyi et al., find the directions affecting age and sex with a custom classifier and for each corresponding latent vector they generate multiple shifts and select the most accurate manipulations. For instance, in the gender swap task, for each synthetically generated face they compute multiple additions of the gender direction, generating corresponding images of both sexes. Gender paired samples are selected according to a trained custom classifier. Selection is based on the highest probability of belonging to a gender, hence the generated pairs are expected to contain only the most polarized gender transitions. Finally, the paired dataset is used to train the Pix2PixHD network to perform gender swap. Distillation for ageing and rejuvenation follows the same aforementioned principles.

For style mixing instead, the authors used the standard crossover technique employed in editing with StyleGAN. Two latent vectors z_1 and z_2 are injected in the input layers of StyleGAN2 to produce intermediate corresponding vectors w_1 and w_2. The intermediate vector w_1, corresponding to the content vector, is used in the first half of the synthesis network, while in the second half (from a crossover point onward) w_1 is substituted with w_2, the style vector. Images corresponding to w_1 and w_2 are also generated, producing a triplet of content, styled and mixed images. Finally, Pix2PixHD is set to accept 6 channels inputs, the concatenated style and content images, producing a corresponding 3 channels mixed image. In this study, we adopt the exact same procedure proposed by the original authors.

For the final part of our work, we extend the concept of distillation by using two trained StyleGANs to generate paired training samples (Fig.2). We focus on the task of image translation with the goal of mapping photo-realistic faces to drawings and vice-versa. To generate targets depicting drawings of faces, we embed the photo-realistic faces generated with StyleGAN2 in a secondary network trained to produce japanese anime characters. Embedding results are evaluated with the structural similarity index (SSIM) computed between pictures generated from StyleGAN2-FFHQ and respective embeddings (Fig. 2.B). The median SSIM is 0.6, and we filtered out results with an index lower than 0.5.

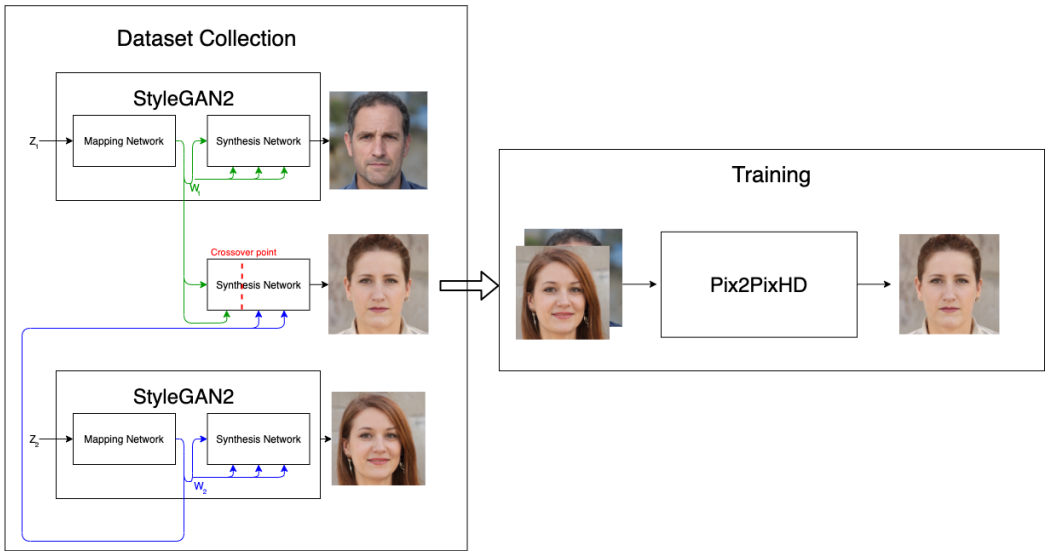

Figure 1: Distillation of StyleGAN2 for the style-mixing task. Two images are generated starting from some normally distributed vectors $z_1 \in \mathbb{R}^{512}$ and $z_2 \in \mathbb{R}^{512}$. The relative w_1 and w_2 vectors are fed in the synthesis network in an alternate manner from a crossover point to obtain the mixed image, which with the former two will constitute a dataset triplet. The dataset is then used to train a vanilla Pix2PixHD network.

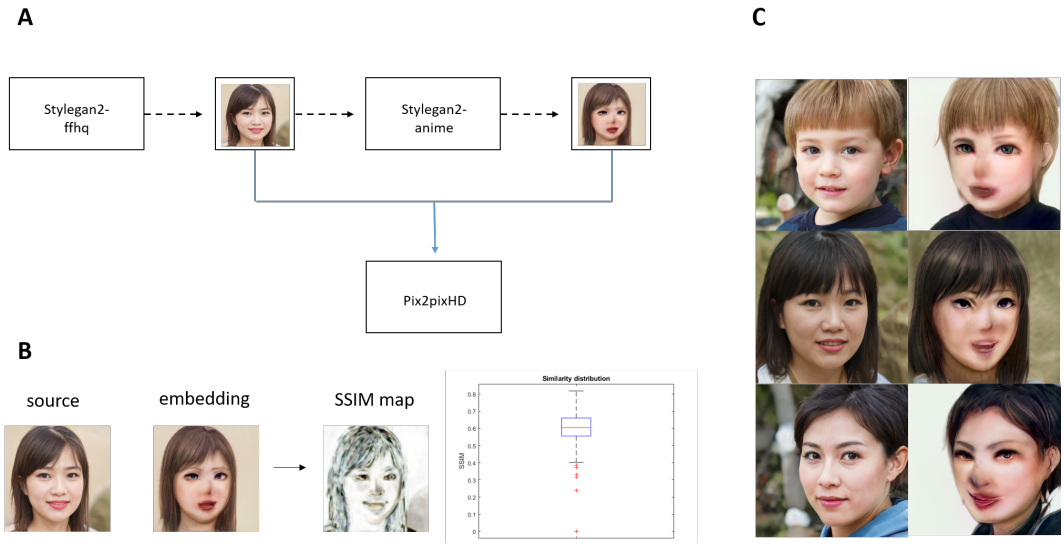

Figure 2: A) Synthetic faces are generated with StyleGAN2 trained on FFHQ dataset. Generated images are embedded with perceptual loss backpropagation in a secondary StyleGAN2 network trained to generate japanese anime characters. B) The quality of the embedded images is evaluated with the structural similarity index (SSIM) using the realistic portraits as references. Embeddings with a SSIM score below 0.6 are discarded. C) Representative gallery of retained embedded images with highest quality scores.

4 Implementation

Our re-implementation study is built on top of two works: StyleGAN2 and Pix2PixHD networks. The StyleGAN2 pipeline to generate our training dataset is based on the original Nvidia repository. For what concerns StyleGAN2, we were able to create a dataset for style-mixing using the original repository and the configuration F of the network trained on the FFHQ dataset. Embedding with backpropagation was performed using the StyleGAN2 encoder repository.

Concerning Pix2PixHD image translation, we could not use the original NVIDIA repository, as it does not provide options to train or test a network on 6-channels input images. This feature is needed as the style and content images are stacked on the channel dimension. Instead, the authors of the original paper provide their own modified version of the Pix2PixHD repository such that the network can be trained on 6-channels input images. We thus make use of their implementation which, to our record, only differs from the original in this regard.

When testing the style-mixing model, input images are first pre-processed as follows: faces are aligned to conform to the FFHQ dataset using the dlib model *shape_predictor_68_face_landmarks*, then, images are converted to RGB and finally reshaped to a size of 256 x 256 pixels. The implementation of the face alignment method is provided entirely by dlib.

4.1 Datasets

To generate synthetic faces we use the configuration F of StyleGAN2 trained on the FFHQ dataset. Random latent vectors of 512 length are sampled from a univariate gaussian distribution with the numpy function randn. Vectors were passed as argument in the generating network Gs . Triplets for style-mixing were generated with intermediate latent vectors crossover, producing content, style and mixed images. Due to cloud services constraints, we reduced the dataset size to 25'000 triplets instead of the 50'000 used in Viazoveskyi et. al. The dataset was generated in ca. 35 minutes on a Google Colab GPU runtime for a total dimension of approximately 10 GB. Although the generated dataset features images with resolution 512 x 512 pixels, the Pix2PixHD network we trained takes as input 256 x 256 x 6 images, as this resulted in a faster training. Each of the training images was successively reshaped to one fourth of the original size.

The dataset employed in face to paint translation was created by embedding generated faces into a secondary StyleGAN2 network trained to generate japanese anime characters. This approach produced paired images of realistic faces and cartoon portrait. Embedding with backpropagation was limited to 1000 epochs, taking approximately 2 minutes per image on a Colab notebook with GPU. The final dataset was composed by 1600 256 x 256 paired images and took approximately 133 hours of discontinuous generation. This dataset was increased with conservative data augmentation introducing only horizontal flipping.

4.2 Hyperparameters

For Pix2PixHD training we mostly used the default hyper parameters . For the The initial learning rate was set to 0.0002 with SGD and momentum equals to 0.5. The batch-size was reduced to 8 to avoid CUDA memory errors while using 64 convolutional filters in the first layer of the network. Lambda, the weight coefficient of the L1 loss in the compound loss metric, is also 10 by default.

4.3 Experimental setup

We run all the experiment on two separate Colab notebooks, providing a single CPU, a NVIDIA Tesla K80 GPU and 16 GB of ram.. Most data manipulation, such as dataset generation from StyleGAN and embedding required a GPU runtime. Hence, we created a shared account Colab premium to run most time consuming processes.

To avoid loosing our training progress due to runtime disconnections, we stored the latest trained generator and discriminator networks in the mounted google drive. Training was successively resumed from the checkpoints several times.

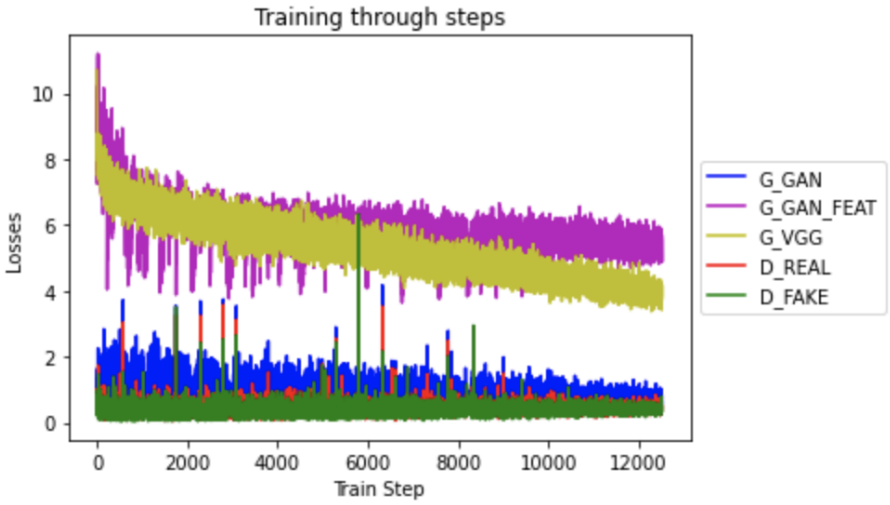

Figure 3: Training losses through steps

Figure 3 shows the default training metrics provided by the Pix2PixHD repository for our style-mixing model over 100 epochs. By the plot, we see that the G_VGG loss seems to be still improvable with epochs, so our final model might be trained even for longer.

The notebook used for our experiments can be found at the following link: `https://colab.research.google.com/drive/1hxZvml_rbjF62W-9bW39Dap1zJgR9K5w?usp=sharing`

`https://colab.research.google.com/drive/1sEI6u8W8UoZvHgR4CqWnXX_lpe-tI1Ge?usp=sharing`

4.4 Computational requirements

We ran our work entirely on Google Colab using a GPU runtime. Our runtime sessions were requiring on average 7 GB of GPU, arriving at 16 GB at most. Our environment was provided with 16 GB of RAM in the free version of Colab, and 32 in the pro version. A large RAM capacity was also required to avoid CUDA memory error during training. The generated trained model were also quite heavy, requiring 600 MB for a generator model with 64 features in the first layer, and peaking at 2.5 GB with heavier models. We recommend thus an environment provided with at least 30 GB of disk.

In terms of time requirements, our style-mixing Pix2PixHD networks were trained for 50 epochs keeping the learning rate fixed, and then for 50 epochs more where the learning rate decayed linearly towards zero. Each epoch took approximately 1'300 seconds or 21.5 minutes to complete. We hence trained the model for a total time of 36 hours and 6 minutes. The models for image translation were less demanding, with an average training session lasting on average 5 hours.

5 Results

For the style-mixing task, we compare our network capabilities with the authors results in mixing fine styles.

(a) Authors results in fine style-mixing task

(b) Our results in fine style-mixing task

Figure 4: Comparison of fine style-mixing results

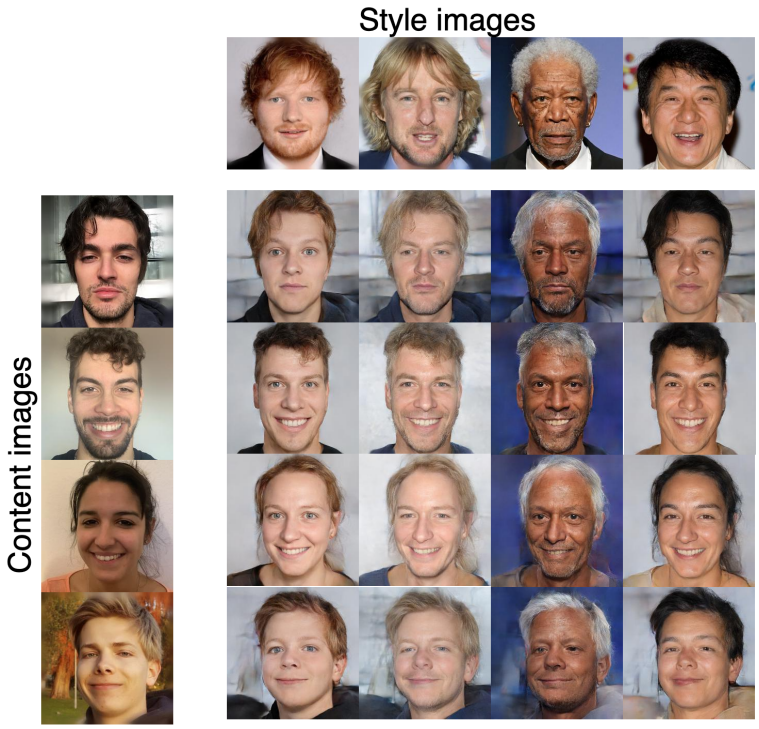

Figure 5: Style mixing with Pix2PixHD model on unseen content and style pictures of real human faces

In figure 4, we see that the model provided by the authors is indeed more capable of producing photo-realistic images. This is to be expected, as the authors trained the model for twice as many epochs with twice as much data. Our model, however, is still capable of producing style mixed images that look photo-realistic in some cases. This can be seen also from figure 5, where the faces of the authors of this paper and friends are mixed with the style of four celebrities. We have noticed that the face alignment between input images is crucial for generating good results as the FFHQ images are processed to be aligned. In Figure 5, the trained Pix2PixHD performs well even on faces that were not part of the FFHQ dataset. This is important to assess to exclude performance biases caused by the training set.

(a) Results in the translation from real faces to drawings.

(b) Results in te conversion of drawings and paintings to realistic faces

Figure 6: Results of Pix2PixHD trained to convert into real faces fine art (left) and stylized paintings(left)

Successively, we tested on real images our Pix2PixHD trained to translate face portraits into drawings (Fig. 6. B). Our testing shows that trained Pix2PixHD underperforms in the face to paint translation task. Despite capturing basic traits such as the skin color, face shape and hairs, the generated images are blurried and similar to an unfinished canvas rather than a proper painting. Particularly, fine details corresponding to eyes, nose and mouth were not successfully captured. Nevertheless, Pix2PixHD generates images which are closer in the feature space to real faces compared to the embedded images. This is possibly due to the network not retaining the artefacts created during the embedding as they are not recognized as patterns.

Surprisingly, we obtained results exceeding our expectations when we trained Pix2PixHD to translate drawings into photo-realistic faces. Despite not capturing the fine details related to a face style, our trained networks reproduced consistently the content of our tested images, face expression included. Our network is clearly immature, producing

different artefacts with increased content degradation associated to more abstract pictures. This is probably due to the limited amount of training sample we used.

6 Discussion and conclusion

In this study, we evaluated the reproducibility of the work of Viazoveskyi et al. with a focus on style-mixing distillation. In their study, the authors worked with 50'000 training samples with 2 GPUs and trained their models for two consecutive days. In our work, to mitigate the lack of computational resources and to speed up the training, we reduced the dimensions of the inputs and outputs of our image-to-image models. Indeed, training a Pix2PixHD network to produce higher resolution images is possible. To this end, the use of a larger synthetic dataset, as well as a longer training time, could be beneficial. Note that this approach was not considered due to computational and storage capabilities.

We assess that the methods proposed in Viazoveskyi et al. is reproducible with satisfactory results. However, it requires extended computational resources, even for a baseline model, which is something that could have been mitigated if a pre-trained model was released. Furthermore, not sharing their custom classifier used for dataset generation and selection strongly undermines the concept or reproducibility. This study shows that improvements in image-to-image translation tasks are possible whenever a GAN model can be used to generate realistic samples.

Concerning the image to paint translation task, we highlighted how it is possible to use a secondary stylized GAN for knowledge distillation. Our approach main limitation is that the StyleGAN2 trained on japanese anime characters was not ideal, generating structurally different embeddings and, therefore, a low quality mapping between realistic and drawed faces. Nevertheless, Pix2PixHD managed to learn basic mappings, such as the shape of the face and hairs, and generalized them to real pictures. It is interesting to note that the network did not learn mappings corresponding to features that in japanese anime characters strongly differs from real humans, such as the nose, the eyes and the mouth. This suggests that Pix2PixHD could have produced better results if we used a secondary network generating images more similar to human faces in the feature space. Using a StyleGAN network trained on fine art paintings may be a proper choice. Nevertheless, Our results are encouraging and may serve as the base for future related works.

References

[1] Angeline Aguinaldo, Ping-Yeh Chiang, Alex Gain, Ameya Patil, Kolten Pearson, and Soheil Feizi. Compressing gans using knowledge distillation, 2019.

[2] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation, 2017.

[3] Victor Besnier, Himalaya Jain, Andrei Bursuc, Matthieu Cord, and Patrick Pérez. This dataset does not exist: training models from generated images, 2019.

[4] Ari Heljakka, Yuxin Hou, Juho Kannala, and Arno Solin. Deep automodulators, 2020.

[5] Zixuan Huang, Jinghuai Zhang, and Jing Liao. Style mixer: Semantic-aware multi-style transfer network. *Computer Graphics Forum*, 38(7):469–480, Oct 2019.

[6] Yuri Viazovetskyi, Vladimir Ivashkin, and Evgeny Kashin. Stylegan2 distillation for feed-forward image manipulation, 2020.

[7] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[8] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2018.

[9] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019.

[10] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan, 2020.