FDSA-GAN: A Frequency-Domain Self-Attention GAN For Improved Line Art Generation Of Anime Faces

Anonymous CVPR submission

Paper ID ****

Abstract

Due to the computational efficiency of GANs during inference and capability for meaningful latent space interpolation, it gained widespread usage among low-compute users for anime face generation, particularly with StyleGAN. While StyleGAN has demonstrated impressive generative capabilities for anime faces, it has a complex architecture built from many empirical observations such as weight demodulation and path length regularization. DC-GAN has a simpler architecture, but has considerably worse generation quality. Inspired by the sparse and high-frequency style of anime line art, we propose FDSA-GAN, which integrates a frequency-domain self-attention block in the early layers of DC-GAN in order to attend to frequency-domain information to capture stronger global frequency understanding for the model to gain better line art generation quality. We use the gochiusa dataset's manga and anime faces. With FDSA-GAN, we observed the most consistent improvement compared to DC-GAN (ranging from 5-10 percent) for Kernel Inception Distance.

1. Introduction

Anime as a cultural medium rose in popularity exponentially in the 21st century. Anime face generation reached a local maximum in popularity during 2019, when people began using StyleGAN on various anime datasets to generate realistic depictions of anime character faces. Anime face generation has numerous use cases, including character sketches for quick prototyping or working as a tool for people not well-versed at drawing and developers who want to design characters with low overhead. Additionally, despite diffusion models tending towards stabler convergence, GANs when tuned well allow for directly meaningful latent space interpolation and more computationally efficient inference (only requiring one forward pass), which is a major benefit for users who have low compute resources.

However, StyleGAN has a complex architecture built from empirical observations, which is difficult for interpretability, which is especially relevant when designing an architectural changes on pre-existing GAN architectures. Currently, DC-GAN, which has a more simple, feasible architecture to build modifications on, doesn't generate as realistic results as StyleGAN. With FDSA-GAN, we add a simple Frequency-Domain Self-Attention block to the early layers of DC-GAN, maintaining the overall simplicity of DC-GAN while improving anime and manga line art generation. Compared to human faces, anime faces have much sharper, high-frequency information along with sparser details or imperfections, making the locality bias of convolutions a potential bottleneck for DC-GAN which can be mitigated with the FDSA block. The input to FDSA-GAN is a random noise vector z from a prior distribution (Gaussian). By passing the noise vector into FDSA-GAN, a generated anime or manga face image, with a focus on accurate line art is outputted.

For training, real anime/manga images are used as data for the discriminator to classify between generated and real images, propagating a signal for the generator to learn from.

Furthermore, we extend analysis towards manga face generation. While GAN-based manga face generation has been less researched compared to generation of RGB anime faces, we believe the black and white nature of manga, which provides more emphasis on line art (our main area of interest), allows manga face generation to be a useful adjacent metric to assess our model's ability to capture high-frequency details and global structure with less color-based noise. We hypothesize that there will be a strong correlation between manga line art and anime line art fidelity.

2. Related Work

2.0.1 Seminal CNN-based GANs (DC-GAN)

While earlier GAN architectures like DC-GAN [11] had success in generating realistic images of human faces (CelebA) or digits (MNIST), they struggled to properly generate anime faces [2]. This was attributed to the general lack of natural, stochastic local textural information of anime faces compared to human faces, which tend to have more inconsistencies/imperfections compared to the clean line art of anime. Considering the nature of early convolutional layers (which have a smaller receptive field), the locality bias is ingrained in typical CNN-based generators, making it difficult to pick up signals in anime faces which focus more on a global context.

2.0.2 Improving GAN Stability, Scalability, and Resolutions

There have been many GAN architectures introduced in the last decade. The Wasserstein GAN (WGAN) [1] used the Lipschitz constraint, Earth Mover's Distance (EMD) with the critic to yield more meaningful losses. BigGAN [3] showed that GANs could be successfully scaled to higher resolution images. ProGAN introduced the concept of progressive growing, such that the model starts with a low spatial dimension and increases with spatial upsampling. [7].

2.0.3 Attention Mechanisms for GANs

Self-attention allows for efficient modeling of long-range dependencies. [13] Though it was originally proposed in the natural language domain, self-attention was extended to visual domain with architectures such as the visual transformer (ViT) [4]. The Self-Attention GAN (SA-GAN) [14] applied the self-attention mechanism to GANs, which was shown to exhibit improved image quality both visually and in metrics like Inception Score (IS) and Frechet Inception distance (FID).

2.0.4 State-of-the-Art Architecture for Fidelity and Style (StyleGAN)

StyleGAN [8] and StyleGAN2 [9] represent state-of-the-art for GAN-based image generation, which extended to not only human faces, but also anime faces and other domains. StyleGAN adopted the progressive growing of resolutions, and also introduced numerous innovations, such as the introduction of a mapping between noise vector z and intermediate vector w (using this w as input for AdaIN), which allowed disentanglement of style. StyleGAN2 used weight (de)modulation as a surrogate for this while mitigating the water droplet problem (artifacts from AdaIN in StyleGAN) and used path length regularization to mitigate the tendency for components to stay fixed during latent space interpolation. Before an input passes through the convolutional layer of a block, its weights are first modulated and demodulated based on the z to w value, independent of the input itself. There have been numerous successful attempts leveraging the use of StyleGAN for anime face generation, including Rem, Emilia, and other anime characters. While StyleGAN has had great results on anime face generation, it used a lot of empirical tricks like minibatch standard deviation that weren't derived from first principles.

2.0.5 Anime-inspired GAN Architectures

The most similar work to FDSA-GAN is USE-CMHSA-GAN [10], which also tries to improve anime character generation, which used upsampling squeeze excitation module [5] with convolution-based self-attention. Meanwhile, our approach focuses on frequency-domain multi-head attention. We start from DC-GAN [11], and build from an observation of anime and especially manga faces: the general lack of mid-frequency information (consider how line art often has wide regions with extremely little variation which can shift to high variation with singular strokes), and surplus of low-frequency and high-frequency information characteristic of line art. This means a main flaw of DC-GAN is in the earlier layers where the simple convolutional blocks do not capture global information well, and do not leverage the information across the frequency domain. Motivated by this observation, we propose a way to improve global frequency-domain contextual understanding in the early layers where the receptive field is still small. We believe this leads to potential improvements in line art generation by adding a frequency-domain self-attention block to gain such information in the earlier layers.

3. Method

General Adversarial Networks (GANs) offer important benefits for generating high-fidelity anime/manga faces, particularly from their capacity for learned latent space interpolation and efficient one-pass inference. Alternative approaches like Diffusion Models require multiple passes during inference for denoising. GANs are capable of implicitly learning high-dimensional data distributions, training a generator network to map noise vectors z from a simple prior distribution (such as Gaussian) to the data manifold. This section details the methodology of our proposed FDSA-GAN. We first briefly note the foundational GAN framework, then compare the architectural design choices of the DC-GAN architecture (baseline) with the StyleGAN architecture (a visual SoTA benchmark). Finally, we introduce our proposed Frequency-Domain Self-Attention (FSDA) module and how it can be integrated into the traditional DC-GAN.

The generator aims to produce realistic fake images, while the discriminator aims to accurately classify between generated and real images. Both the standard DC-GAN and our proposed FDSA-GAN are trained using the standard GAN minimax loss formulation as follows:

$$\begin{aligned}\min_G \max_D V(D,G) &= \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] \\ &+\mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))]\end{aligned} \tag{1}$$

3.1. DC-GAN

The baseline architecture, DC-GAN [11], was a seminal architecture that improved GAN-based image generation. It highlighted a number of important design choices, such as the deep convolutional structure, replacement of pooling layers with strided and transposed convolutions, and use of batch normalization, which allowed for stabler training and improved image generation quality. DC-GAN defines both Generator G and Discriminator D as deep Convolutional Neural Networks. DC-GAN replaces pooling layers with strided convolutions for the discriminator, and with transposed convolutions for the generator. It uses the ReLU activation for the generator, and LeakyReLU for the discriminator, with batch normalization for both architectures.

3.1.1 DC-GAN Generator Architecture:

Given a latent vector $z \in \mathbb{R}^{d_z}$, the generator $G(z; \theta_g)$ maps z to image $G(z) \in \mathbb{R}^{C \times H \times W}$. The generator consists of a series of L upsampling blocks. For $l = 0, \ldots, L-1$,

$$\begin{aligned} h^{(l+1)} &= G^{(l+1)}(h^{(l)}; \theta_g^{(l+1)}) \\ &= \text{ReLU}(\text{BN}(\text{ConvTranspose}(h^{(l)}))) \end{aligned} \tag{2}$$

where BN represents batch normalization and ConvTranspose is a 2-dimensional transposed convolution.

3.1.2 DC-GAN Discriminator Architecture:

The discriminator follows a similar structure to the generator. $D(x; \theta_d)$ maps $x \in \mathbb{R}^{C \times H \times W}$ to scalar. Assuming a series of L stacked layers, for $l = 1, \ldots, L-1$:

$$\begin{aligned} h^{(l+1)} &= D^{(l+1)}(h^{(l)}; \theta_d^{(l+1)}) \\ &= \text{LeakyReLU}(\text{BN}(\text{Conv2d}(h^{(l)}; \theta_d^{(l+1)}))) \end{aligned} \tag{3}$$

As the final discriminator (output) layer, the sigmoid activation is applied to convert logits to probability.

$$D(x; \theta_d) = \sigma(output) \tag{4}$$

3.2. StyleGAN

In contrast with the DC-GAN's simple architecture, StyleGAN has a large number of architectural changes, motivated by empirical observations. StyleGAN achieved high-resolution image quality, along with disentangled latent space and style control. StyleGAN's success in generating not only high quality human faces, but also the first GAN architecture to generate accurate anime faces, makes it a strong SoTA qualitative benchmark for evaluating new GAN architectures for anime/manga face generation.

Let $f(\cdot; \phi_f)$ be the mapping network, $g(\cdot; \theta_g)$ the synthesis network. $G(z; \phi_f, \theta_g) = g(f(z; \phi_f); \theta_g)$.

3.2.1 Mapping Network (f):

In order to promote disentanglement, z is mapped through an 8-layer MLP to yield w, a more disentangled intermediate vector. It transforms $z \in \mathcal{Z}$ to $w \in \mathcal{W}$ as follows:

$$w = f(z; \phi) \tag{5}$$

3.2.2 Adaptive Instance Norm (AdaIN) and (De)Modulation:

With layer i of g, the transformation $MLP_i(\cdot; \phi_i)$ maps w to styles $(y_{s,i}, y_{b,i})$ with s denoting scale and b denoting bias:

$$(y_{s,i}, y_{b,i}) = MLP_i(w; \phi_i) \tag{6}$$

Defining $x_k^{(i)}$ as the k-th feature channel at layer i, the AdaIN operation can be described as follows:

$$\begin{aligned} \text{AdaIN}(x_k^{(i)}, y_{s,k}^{(i)}, y_{b,k}^{(i)}) = y_{s,k}^{(i)} \frac{x_k^{(i)} - \mu(x_k^{(i)})}{\sqrt{\sigma(x_k^{(i)})}} \\ + y_{b,k}^{(i)} \end{aligned} \tag{7}$$

The purpose of AdaIN was to use w was to normalize the feature activations and apply a specific scale and bias. The style/intermediate vector w is used to modulate the normalized features in order to inject a distinct style. However, it was noted that normalization produced ”water droplets” [9]. In StyleGAN2, AdaIN was replaced by modulation and demodulation. There are also many additional architectural choices, such as noise injection and path length regularization which help StyleGAN achieve maximum output quality.

3.3. Frequency-Domain Self-Attention (FDSA)

3.3.1 Motivation

The design of the FDSA block is motivated by the receptive field limitations of early convolutional layers for the generator. The earlier layers of DC-GAN consist of simple convolutional blocks which are not global frequency-aware. Standard convolutions can gain larger receptive fields in deeper layers, but earlier layers typically do not have direct access to the full tensor. For anime faces, which has much sparser local information compared to human faces, the representation capabilities of standard convolutional layers may be lower, making it more difficult to capture global context. Additionally, the line art quality of anime faces is much sharper than human faces, making the information from anime faces higher frequency. The FDSA block aims to address these limitations by introducing a self-attention block in the frequency domain to gain such information in the earlier layers, even in the earlier layers the GAN should be able to gain a stronger global representation and understand high-frequency quality of line art to generate more accurate anime faces.

3.3.2 Baseline Implementation Details

The Frequency-Domain Self-Attention block will be integrated into this baseline DC-GAN, which follows the architectural details outlined in DC-GAN. We chose to implement DC-GAN from scratch with pytorch given the objective of integrating an efficient, simple architectural change to DC-GAN to get closer performance to StyleGAN's complex architecture, particularly in image fidelity focused on the line art style of anime. The original DC-GAN generator starts with a mapping from noise vector z to a tensor (1024, 4, 4). It then consists of spatial upsampling blocks (with channel downsampling), using transposed convolutions for upsampling, then applying batch normalization with ReLU. The DC-GAN discriminator consists of downsampling blocks using strided convolutions, with batch normalization and LeakyReLU activation.

The successful anime generation model I will compare my improved model against is StyleGAN2 with WGAN-GP loss function. For StyleGAN, we used an already implemented version from LabML [6], which uses an MLP to map the z noise vector to an intermediate vector and usage of weight modulation and demodulation. WGAN improved the training stability of GANs along with a meaningful loss by adopting Wassserstein distance, Lipschitz constraint, and the critic.

3.3.3 FDSA Overview

The Frequency-Domain Self-Attention (FDSA) block is inserted right after the projection of noise vector z, before convolutional layers, operating on the dimension 1024 x 4 x 4. The early placement of the FDSA block is chosen for two reasons: smaller spatial dimensions reduce the computation costs of the quadratic self-attention mechanism on the frequency domain, and increasing global contextual understanding in the earlier layers can propogate to deeper layers, which can augment their larger receptive fields. In situations with moderately high compute, the FDSA block could be introduced in later layers (such as 8x8 or 16x16) to obtain more frequency bins for more fine-grained global context. Continuing from the pre-conv layer placement assumption, A 2D FFT is applied to each of the 1024 channels of the 4x4 feature maps, yielding complex tensors. Each spatial location in the frequency-domain feature map forms a token along the channels, yielding 16 spatial frequency locations (tokens). After concatenation of real and imaginary components, each token will have dimension 2048 (1024 x 2). Given these tokens, we can apply standard self-attention [13]

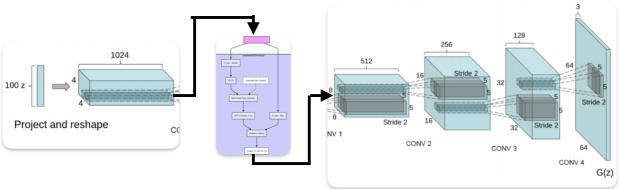

Figure 1. DCGAN + FDSA Generator

We first want to transform the input X_{in} to frequency domain using FFT2D to get $X_{\text{complex\_freq}} \in \mathbb{C}^{B\times C\times H\times W}$. Then we split $X_{\text{complex\_freq}}$ into separate real and imaginary parts, also concatenating them along the channel dimension:

$$X_{\text{real\_part}} = \text{Re}(X_{\text{complex\_freq}}) \tag{8}$$

$$X_{\text{imag\_part}} = \text{Im}(X_{\text{complex\_freq}}) \tag{9}$$

$$X_{\text{freq}} = \text{Concat}([X_{\text{real\_part}}, X_{\text{imag\_part}}]) \tag{10}$$

This yields $X_{\text{freq}} \in \mathbb{R}^{B\times C'\times H\times W}$, such that $C' = 2C$. X_{freq} is reshaped into $X_{\text{tokens}} \in \mathbb{R}^{B\times N\times C'_{\text{embed}}}$, where each spatial location in the frequency-domain feature map forms a token along the channels, making there be H x W tokens of dimension C'_{embed}. For simplicity, let $C'_{\text{embed}} = C'$. Now we can use traditional multi-head self-attention on these tokens.

3.4. Multi-Head Self-Attention (MHSA)

Let N_h be number of heads, $d_k = C'_{\text{embed}}/N_h$. Following the Pre-LayerNorm concept, Layer Normalization is applied to X_{tokens} before the QKV projections. For head i, with $W_{Q_i}, W_{K_i}, W_{V_i} \in \mathbb{R}^{C'_{\text{embed}}\times d_k}$:

$$Q_i = \text{LN}(X_{\text{tokens}})W_{Q_i} \tag{11}$$

$$K_i = \text{LN}(X_{\text{tokens}})W_{K_i} \tag{12}$$

$$V_i = \text{LN}(X_{\text{tokens}})W_{V_i} \tag{13}$$

such that $Q_i, K_i, V_i \in \mathbb{R}^{B\times N\times d_k}$. Now we can perform the typical self-attention calculation.

$$H_i = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i \tag{14}$$

Then concatenate all H_i to $H_{\text{concat}} \in \mathbb{R}^{B\times N\times C'_{\text{embed}}}$ and perform the output projection:

$$X_{\text{att}} = H_{\text{concat}} \cdot W_O \tag{15}$$

3.5. Residual Connection and Normalization

Now, X_{att} is reshaped, to split into the real and complex components, and obtain the output complex frequencies such that the 2D IFFT can be applied, in which we can obtain the real component output $X_{\text{spatial\_att}}$ such that X_{in} represents the original spatial domain input tensor (pre-FDSA block)

$$X_{\text{out}} = \text{LN}(X_{\text{in}} + X_{\text{spatial\_att}}) \tag{16}$$

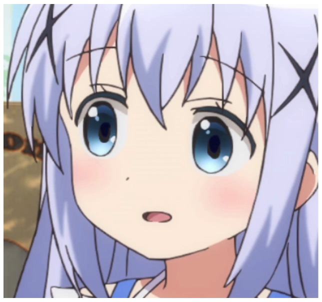

Figure 2. Example Anime Image

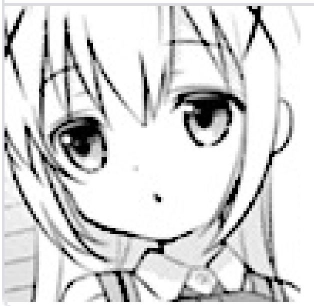

Figure 3. Example Manga Image

4. Dataset and Features

The gochiusa dataset [12] was used to train FDSA-GAN, which consisted of 39537 images from the anime *Is This Order A Rabbit*. We used a 80-10-10 train-validation-test split. Importantly, the resolutions of these images varied largely, from 26x26 to 987x987, with a 356x356 median resolution. In our experiments, we directly resized all images to 64x64, which was the input size FDSA-GAN trained on due to the low compute resources available. In future research, the impact of training with different original image resolutions will be experimented with by dividing the dataset into general resolution bins. Since FDSA-GAN is built from a DC-GAN base, we directly use the input images in their resized dimensions.

Example images are shown in Figures 2 and 3.

5. Experiments

Most of the hyperparameters follow the methodology of DC-GAN, besides the integration of the FDSA block. We use the Adam optimizer, and experiment with two learning rates, specifically lr=0.0002 and lr=0.001. We initialize $\beta_1 = 0.5$ and $\beta_2 = 0.999$. We also experimented with different dimensions of the latent vector z, specifically dimension 50, 100, and 200. FDSA-GAN also uses the same number of layers as DC-GAN as shown on Figure 1. Training was conducted with a batch size of 64 images.

The most important method of evaluating the effectiveness of FDSA-GAN is through visual comparisons to DC-GAN and StyleGAN. The expected results are that by introducing the frequency domain self-attention block, the anime face cohesiveness will be integrated with line art style better than the vanilla DC-GAN. Specifically, qualitative evaluation includes viewing the general sharpness of the images, specifically the line art quality of the anime faces and hair style. An ideal anime face should be drawn sharply, without the blob-like structure which GAN outputs often exhibit for anime faces. Additionally, we introduce the concept of comparing anime face generations from manga images vs anime images, expecting that manga images due to the black-and-white coloring should prove easier for most GAN architectures to mimic, which could serve as another baseline that distinguishes model improvements from a general understanding of line form in monochrome conditions compared to more noisy, colorful conditions.

Quantitatively we evaluate the performance of FDSA-GAN using numerous metrics. We compare the distributional similarity between generated and real images with Kernel Inception Distance (KID) for FDSA-GAN vs DC-GAN. This score has significance, but the task of exhibiting sharper line art is something that can be found by visual inspection quite easily, so the qualitative evaluation is more directly important. No cross-validation was done. Due to the higher computational costs for training GANs compared to classification CNNs, we focused primarily on anime face generation, which has a larger target audience, but also perform a quantitative comparison between DC-GAN and FDSA-GAN for the manga dataset.

We also consider other metrics, such as the Deep Image Structure and Texture Similarity (DISTS), Learned Perceptual Image Patch Similarity (LPIPS), and SSIM (Structural Similarity Index Measure) between FDSA-GAN and DC-GAN. A lower Deep Image Structure and Texture Similarity score indicates more structural similarity such that the learned deep features and generated and real images are more similar. A lower Learned Perceptual Image Patch Similarity score, using deep features from pre-trained CNNs, indicates better perceptual similarity, which is more similar to human preferences.

5.0.1 StyleGAN Qualitative Analysis

I first generated anime images with StyleGAN. The outputs demonstrated extremely sharp hair quality characteristic of anime, as well as mostly well-formed facial structure. This can be seen from the first row of Figure 4. While there were some flawed image generations in the second row, particularly the face at coordinate (2,3), the overall line art quality was still noticeably superior to the DC-GAN anime face generations. The middle image also has issues around the chin, where it seems like a chunk of the face is missing. Overall though, the quality is quite high.

Table 1. Quantitative Results for Anime Face Generation.

Model Configuration	KID	DISTS	LPIPS	SSIM
DCGAN (z=100, lr=2e-4)	0.22 ± 0.01	0.332 ± 0.006	0.48 ± 0.01	0.187 ± 0.006
FDSA-GAN (z=100, lr=2e-4)	0.21 ± 0.01	0.301 ± 0.005	0.47 ± 0.01	0.211 ± 0.004
DCGAN (z=200, lr=2e-4)	0.23 ± 0.01	0.312 ± 0.005	0.48 ± 0.01	0.192 ± 0.004
FDSA-GAN (z=200, lr=2e-4)	0.21 ± 0.01	0.303 ± 0.005	0.47 ± 0.01	0.192 ± 0.004
DCGAN (z=50, lr=2e-4)	0.19 ± 0.01	0.313 ± 0.006	0.47 ± 0.01	0.222 ± 0.004
FDSA-GAN (z=50, lr=2e-4)	0.19 ± 0.01	0.303 ± 0.005	0.46 ± 0.01	0.212 ± 0.005
DCGAN (z=100, lr=1e-3)	0.20 ± 0.01	0.331 ± 0.005	0.53 ± 0.01	0.172 ± 0.005
FDSA-GAN (z=100, lr=1e-3)	0.18 ± 0.01	0.329 ± 0.004	0.52 ± 0.01	0.191 ± 0.007

Table 2. Quantitative Results for Manga Face Generation.

Model Configuration	KID	DISTS	LPIPS	SSIM
DCGAN (z=100, lr=0.0002)	0.304 ± 0.002	0.304 ± 0.006	0.444 ± 0.009	0.129 ± 0.006
FDSA-GAN (z=100, lr=0.0002)	0.299 ± 0.001	0.294 ± 0.003	0.436 ± 0.004	0.130 ± 0.006

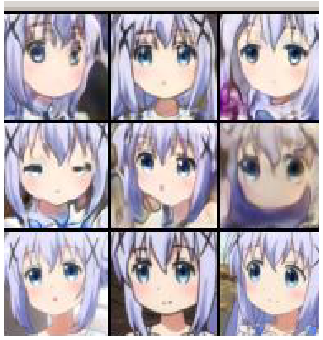

Figure 4. StyleGAN for Anime Images

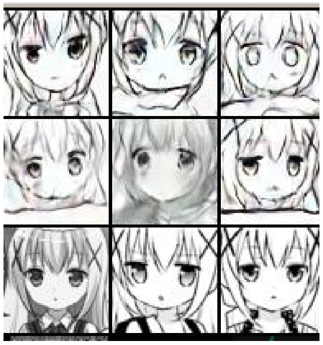

Figure 5. StyleGAN for Manga Images

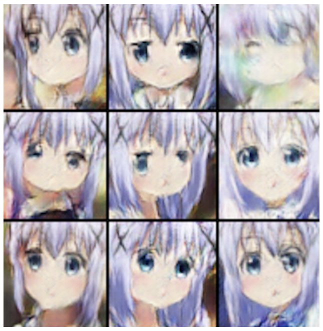

Figure 6. DC-GAN for Anime Images

I then generated manga images with StyleGAN. Despite the greater emphasis on line art due to the monochrome quality of manga, StyleGAN seemed to reach a similar level of quality in its image generations compared to anime. In Figure 5, StyleGAN also had poor image generations, particularly with the images on the left which had deformed facial structure. Interestingly, the top left image exhibited a mostly sharp deformation in the facial structure, while the middle left image showed a more blob-like deformation. This points to the presence of two types of generation errors, one from general facial structure, and the other from lack of frequency information. Overall, with the manga face generations, there was also more residue for some of the images which indicated a lack of full understanding towards the line art style that defines manga.

5.0.2 DC-GAN vs FDSA-GAN Qualitative Analysis (Anime)

Next, I tried evaluating the anime outputs from DC-GAN. As expected, in Figure 6 the vanilla DC-GAN had major issues with anime face generation, with numerous generated faces showing eye deformations, face deformations, and fading hair color. For FDSA-GAN as shown in Figure 7, while face deformation still occurred as shown in the bottom right image, the degree was not as large as DC-GAN, and the occurrences were less. Additionally, there were no cases where the hair color had the fading effect, or the entire image was deformed. Generally, there was clear qualitative

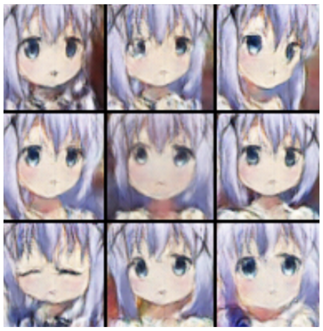

Figure 7. DC-GAN + FDSA for Anime Images

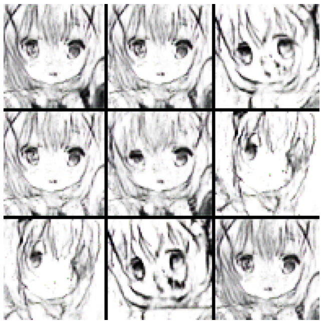

Figure 8. DC-GAN for Manga Images

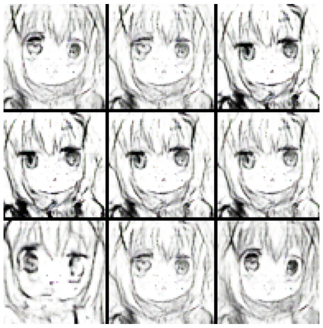

Figure 9. DC-GAN + FDSA for Manga Images

improvement in anime face generation, and the generation quality of FDSA-GAN was much closer to StyleGAN's outputs, though StyleGAN still had improvements in general clarity.

5.0.3 DC-GAN vs FDSA-GAN Qualitative Analysis (Manga)

Finally, I tried evaluating the manga outputs from DC-GAN. Again as expected, the vanilla DC-GAN is far worse at understanding the line art of manga compared to StyleGAN or FDSA-GAN. The issues seemed to grow even more apparent, with 4/9 images in Figure 8 exhibiting glaring issues in the generation quality, with random artifacts occurring on the face, the hair. For the bottom left example, one of the eyes even seems completely missing, showing the inability for DC-GAN to learn the line art style of manga even more so than for anime. In contrast, FDSA-GAN showed much better generation quality, with only 1/9 images in Figure 9 showing issues in generation quality. Overall, it seems that with manga face generation, the use of the Frequency-Domain Self-attention block becomes more important, potentially due to the even greater lack of local details for the network to learn from, furthering the importance of understanding high-frequency information to adjust for the increased importance of line art understanding.

5.0.4 DC-GAN vs FDSA-GAN Quantitative Analysis

Quantitatively, we experimented with a variety of z dimensions (50, 100, 200) and learning rates (2e-4, 1e-3). We observed that the metrics were mostly consistent across anime and manga, such that FDSA-GAN improved KID metric. Besides in the case (z=50, lr=2e-4) in which the Kernel Inception Distance was the same between DC-GAN and FDSA-GAN, we observed that FDSA-GAN had 5-10 percent improvements in KID. It is possible that due to the smaller z dimension, the image generation task was simpler both for DC-GAN and FDSA-GAN due to the generator needing to learn a simpler mapping from the latent vector to image. Based on these observations, FDSA-GAN tends to have great improvements in image quality when the latent vector is of higher dimension, due to how a higher dimensional latent space has greater capacity for encoding image variations.

For the other metrics, we observed that the both DISTS and LPIPS scores had slight improvements when using FDSA-GAN. For DISTS, FDSA-GAN consistently reached lower scores, while for LPIPS, the effects were more marginal for the anime dataset. However, for the manga dataset, the LPIPS score had a larger improvement, suggesting that FDSA-GAN was more effective for improving the LPIPS score in the case with manga (less local color-based noise). For SSIM, the results were very mixed, indicating that the luminance and contrast of generated anime and manga images for FDSA-GAN and DC-GAN were mostly the same. This implies that while the FDSA block had marked improvements regarding the overall facial structure, the frequency-domain self-attention wasn't very effective in improving qualities like luminance or contrast.

6. Conclusion

Our proposed FDSA-GAN, which integrates a frequency-domain self-attention block in the early layers of DC-GAN in order to better attend to frequency-domain information to gain global frequency statistics, in order for the model to gain better line art generation quality. The integration of the FDSA block to DC-GAN maintains the general structure of DC-GAN, but enables higher-quality line art generation both for anime and manga faces. It achieves a middle ground between the complex architecture but SoTA generation results of StyleGAN, the simpler architecture but considerably worse generation quality of DC-GAN. The primary metrics which FDSA-GAN exhibited improvements ranging from 5-10 percent compared to DC-GAN was Kernel Inception Distance. For future work, given higher compute we would explore the effects of binning different resolution images into separate datasets, and scale up FDSA-GAN beyond 64x64 images. We would also experiment with different placement locations of the FDSA block to test for optimal placement configurations.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan, 2017. 2

[2] G. Branwen. Making anime faces with stylegan, 2019. 1

[3] A. Brock, J. Donahue, and K. Simonyan. Large scale gan training for high fidelity natural image synthesis, 2019. 2

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 2

[5] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu. Squeeze-and-excitation networks, 2019. 2

[6] V. Jayasiri. labml.ai: A library to organize machine learning experiments, 2020. 4

[7] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation, 2018. 2

[8] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks, 2019. 2

[9] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan, 2020. 2, 3

[10] J. Lu. Enhanced anime image generation using use-cmhsa-gan, 2024. 2

[11] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016. 1, 2, 3

[12] rignak. Gochiusa faces dataset. 5

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2023. 2, 4

[14] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks, 2019. 2