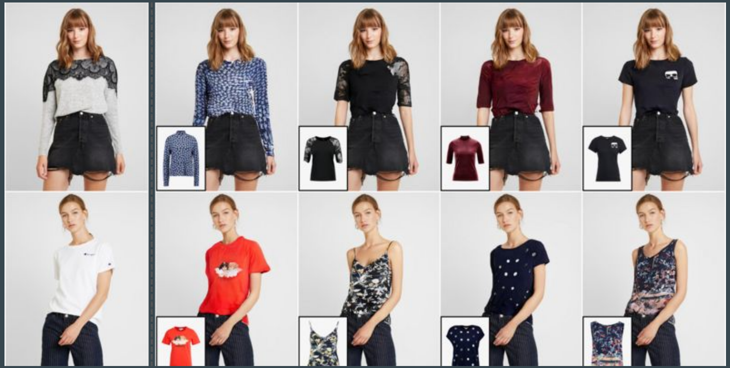
Comparing GAN Architectures Using Fashion Image Generation on Zalando's Viton-HD



Goal: As easily as possible, generate okay images for comparison of baseline models with minimal training.

Options:

- DCGAN (Deep Convolutional GAN)
- WGAN-GP (Wasserstein GAN with Gradient Penalty)
- InfoGAN
- VAE-GAN (Variational Autoencoder + GAN)
- StyleGAN (Lite)

Setup

- 11,647 images
- resized to 64x64
- All GANs trained for 30 epochs
- Batch size 64
- masks overlaid on images to get white background grey to distinct from clothing.

Evaluation metric

1000 generated images

- Fréchet Inception Distance (FID)

compares the **feature distributions** of real and generated images using statistics of activations from a pre-trained Inception network.Computes **mean** and **covariance** of real and generated image features. Measures the **Fréchet distance** between these distributions. Assumes features are Gaussian-distributed. Biased for small sample sizes (<10,000 images).

- Kernel Inception Distance (KID)

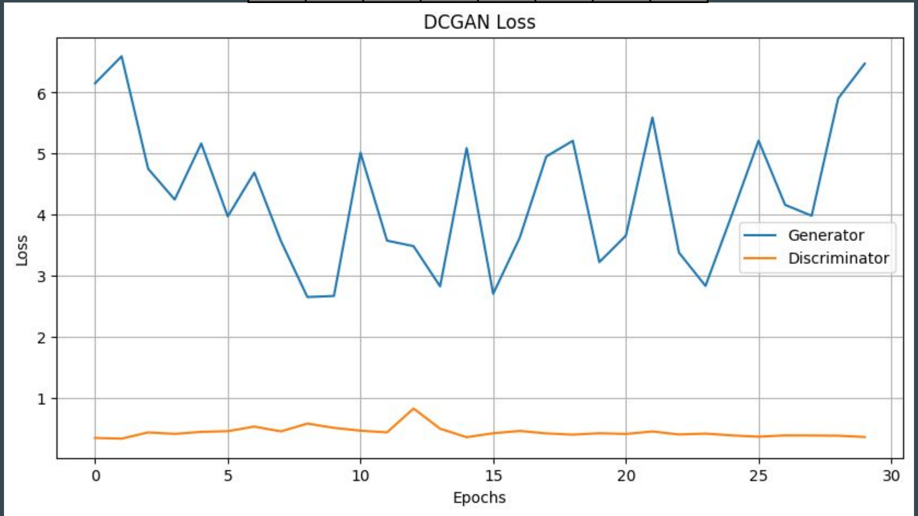
compares features extracted from an Inception network, but uses the **Maximum Mean Discrepancy (MMD)** between them.**Unbiased** even for small sample sizes (ideal for 1000-2000 images).
Better behaved than FID when comparing models on limited data.
Doesn’t assume a Gaussian distribution of features. Computes MMD between real and fake features using a polynomial kernel. Outputs a similarity score + variance estimate.

Deep Convolutional GAN

FID Score: generated_dcgan 302.6959261690736

KID (generated_dcgan): 0.3296 ± 0.0186

Generator loss is high and fluctuating.
Discriminator loss is very low and flat → **D is overpowering G**.



Generator

- Layers: 8 (Linear + 4 ConvTranspose2d + BatchNorm + Activations)
- Parameters: ~6.3 million

Discriminator

- Layers: 7 (4 Conv + Linear + BatchNorm + Activations)
- Parameters: ~2.8 million

Wasserstein GAN with Gradient Penalty

FID Score: generated_wgan 262.95251622304636

KID (generated_wgan): 0.2705 ± 0.0161

Losses are negative (expected for WGAN), but G is falling sharply while D stabilizes.
Looks healthy; G is improving while D becomes consistent.

Same generator as DCGAN

Discriminator

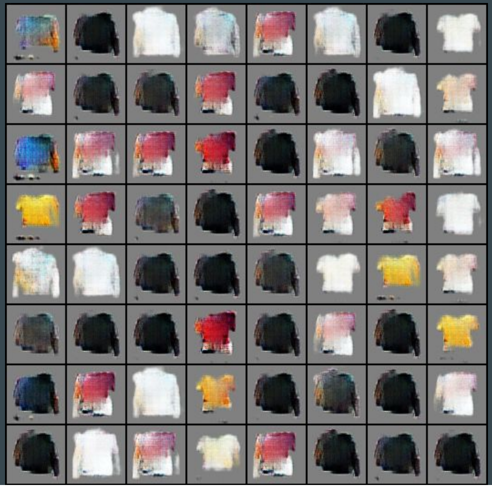
- Layers: Similar to DCGAN, but with **SpectralNorm** and **no Sigmoid**
- Parameters: ~2.8 million

Info GAN

FID Score: generated_info 330.8660912431193

KID (generated_info): 0.3343 ± 0.0144

G loss is slowly climbing, D loss is erratic → **training instability**.
Possibly undertrained Q-network or poor conditioning.



Generator

- Layers: 8 (Linear + ConvTranspose2d + BatchNorm + Activations)
- Parameters: ~1.4 million

Discriminator

- Layers: 2 convs + Flatten + 2 linear heads (real/fake + Q head)
- Parameters: ~1.3 million

VAE-GAN

FID Score: generated_vae 269.294765470754

KID (generated_vae): 0.2924 ± 0.0141

Generator loss stays high with big spikes.
Discriminator flat → might not be learning much.

Encoder: 3 Conv + Flatten

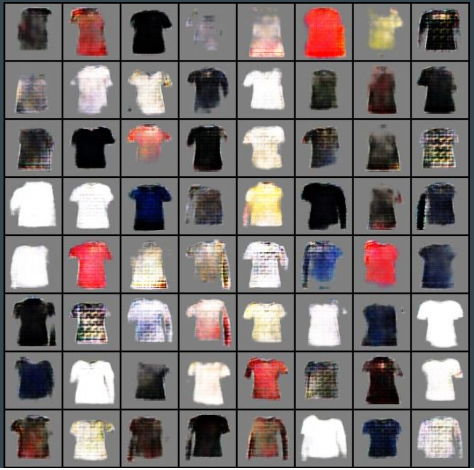
- Parameters: ~2.3 million

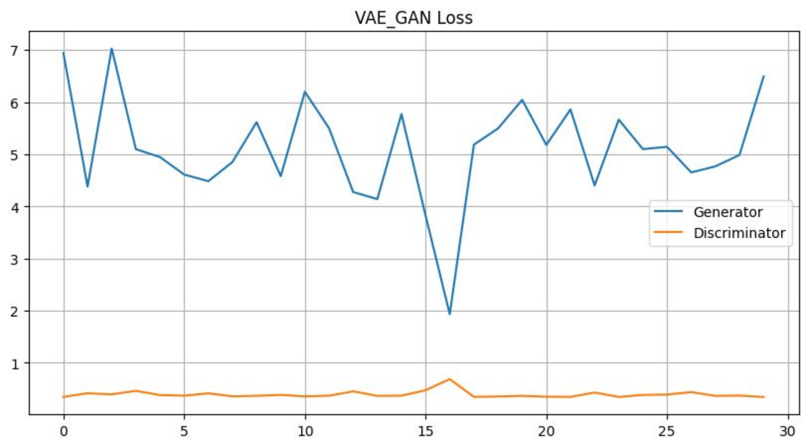
Decoder: Linear + 3 ConvTranspose

- Parameters: ~1.8 million

Discriminator: 3 Conv + Flatten

- Parameters: ~1.1 million





Lite StyleGAN

FID Score: generated_style 300.15344082243837

KID (generated_style): 0.2864 ± 0.0159

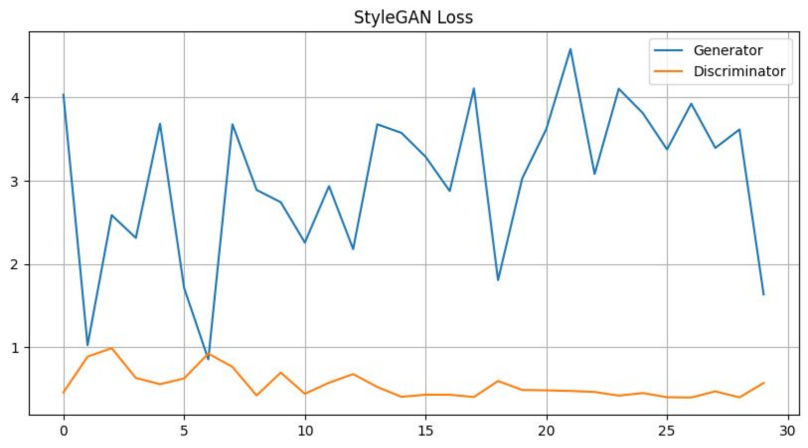
G and D losses are closer, more dynamic interaction.
Slight oscillation is **normal for adversarial training**.

Generator

- Layers: Mapping (2 Linear) + 4 Upsample blocks (each: ConvTranspose + AdaIN + Activation)
- Parameters: ~2.0 million

Discriminator

- Layers: 3 Conv + Adaptive Pool + Linear
- Parameters: ~0.9 million



Results

Model	FID	KID
DCGAN	302	0.3296
WGAN	**262**	**0.2705**
InfoGAN	330	0.3343
VAE-GAN	269	0.2924
Lite StyleGAN	300	0.2864

InfoGAN struggles the most - possibly due to the added conditioning space c being underutilized or hard to learn.

Limitations

- Google Colab Pro+ or dedicated GPU
- Increased Resolution of images
- More epochs
- Better tweaking of parameters and layers

InfoGAN is harder to train because it must simultaneously learn meaningful latent codes and generate realistic images, which requires careful balance between mutual information and adversarial objectives.

VAE-GAN is difficult because the VAE’s goal of accurate reconstruction conflicts with the GAN’s push for realism, making it challenging to optimize both the encoder-decoder and the discriminator together.