Learning Music

STYLE TRANSFORMATION

HONGBO ZHU

STYLE A

eg: Country music
Violin
Man voice

STYLE B

eg: classic music
Piano
Women voice

Cycle generative adversary network

MAN

G

WOMEN

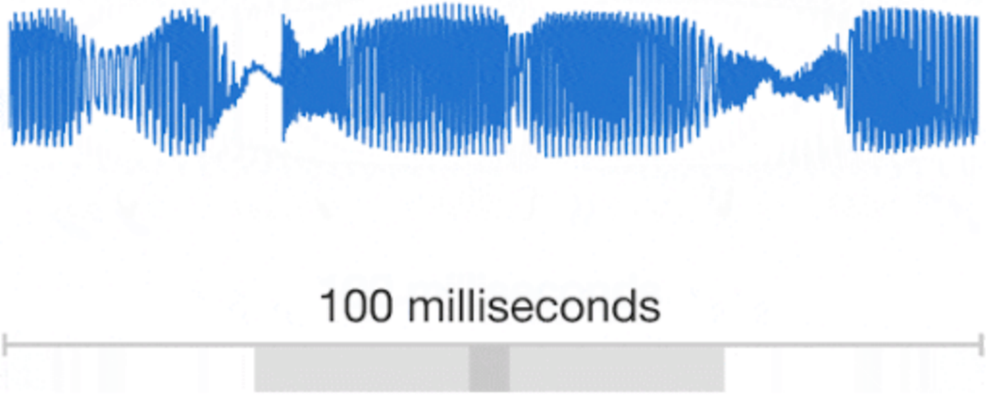
F

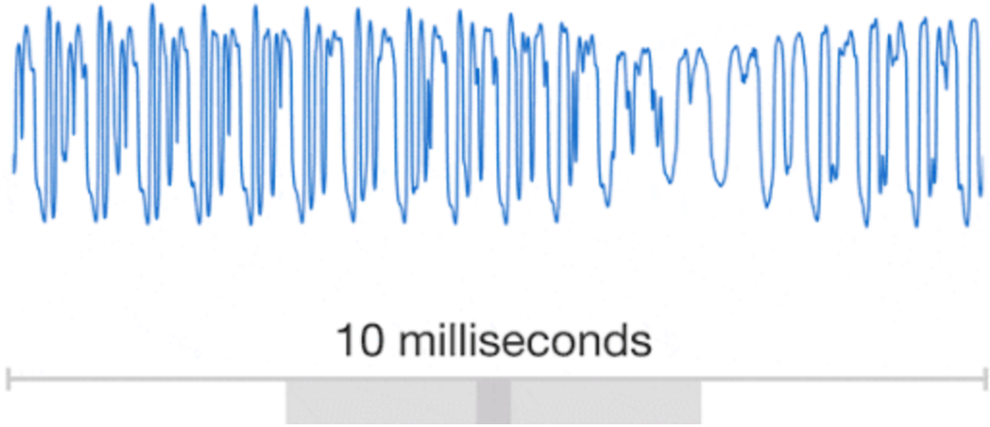
D_A

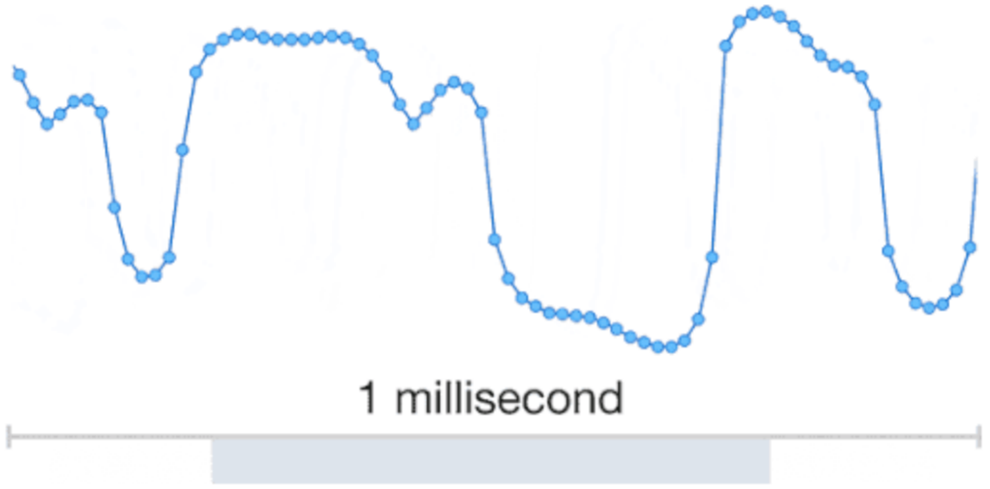
D_B

HOW TO DEAL WITH MUSIC DATA?

Music data representation

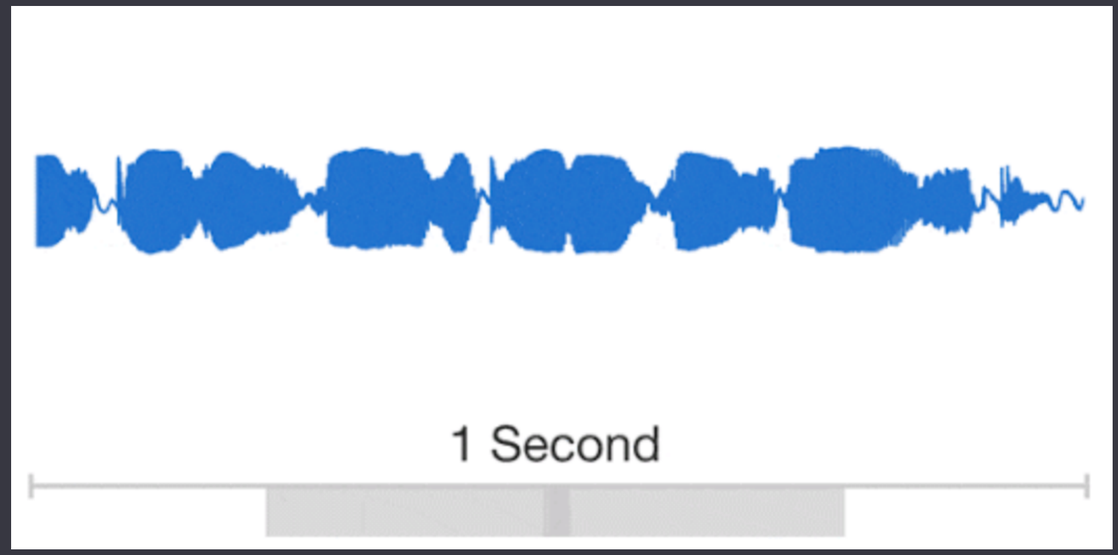


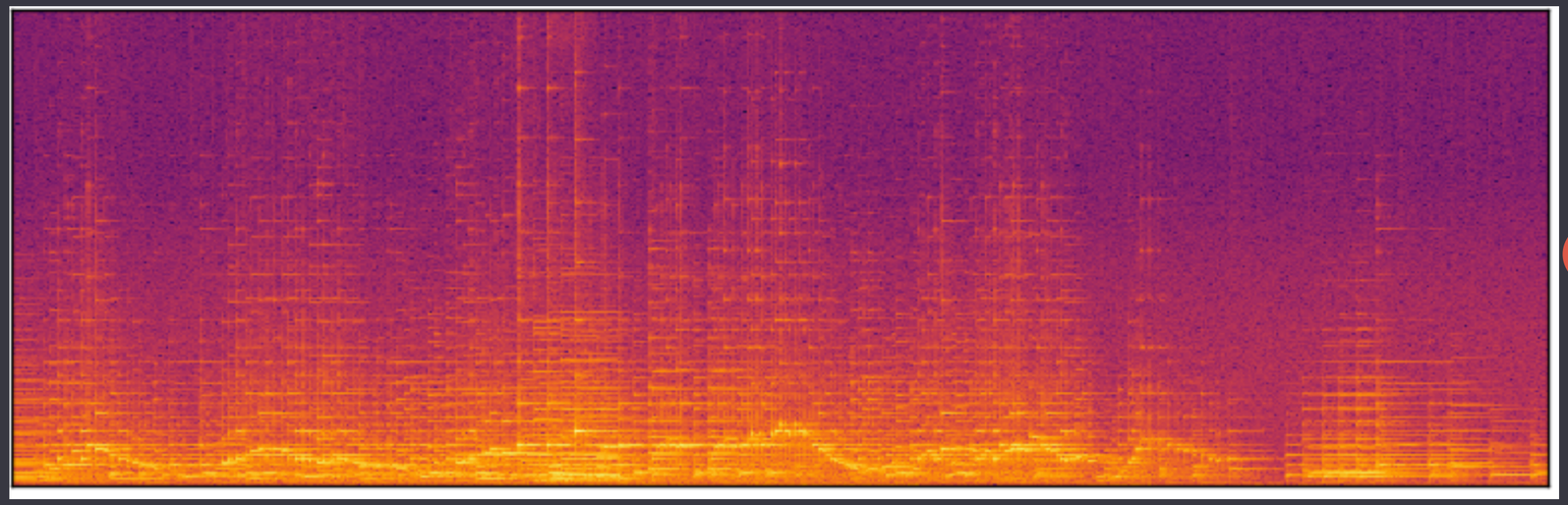




Spectrogram-STFT

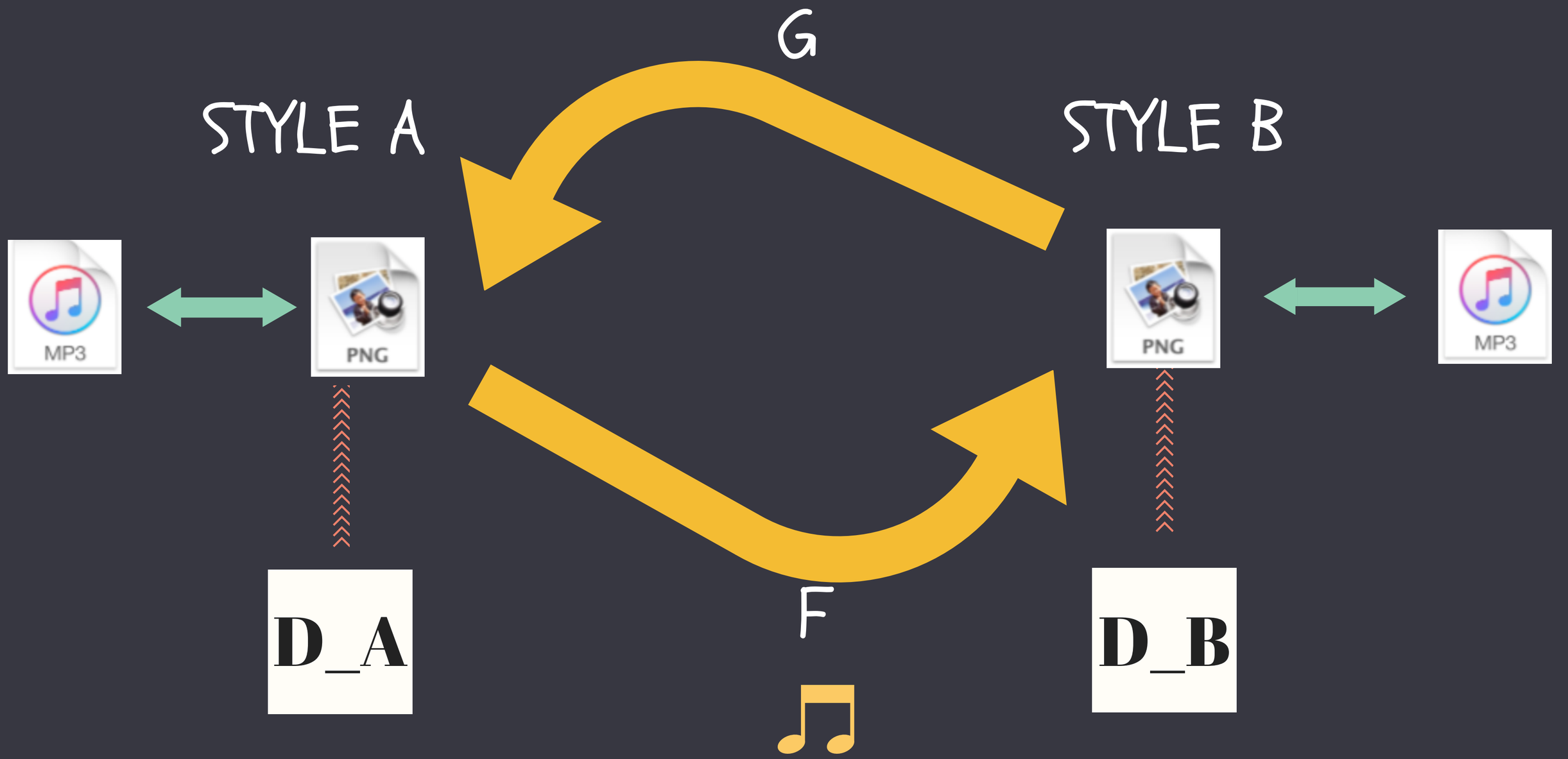
- Duration 30s
- Frame-size 30ms
- Frame-step 10ms
- FFT-size 512
- Audio sampling rate of 20.5kHz.



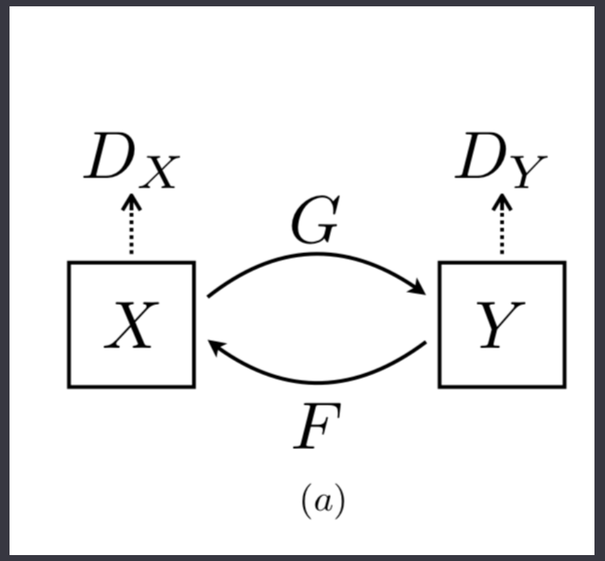


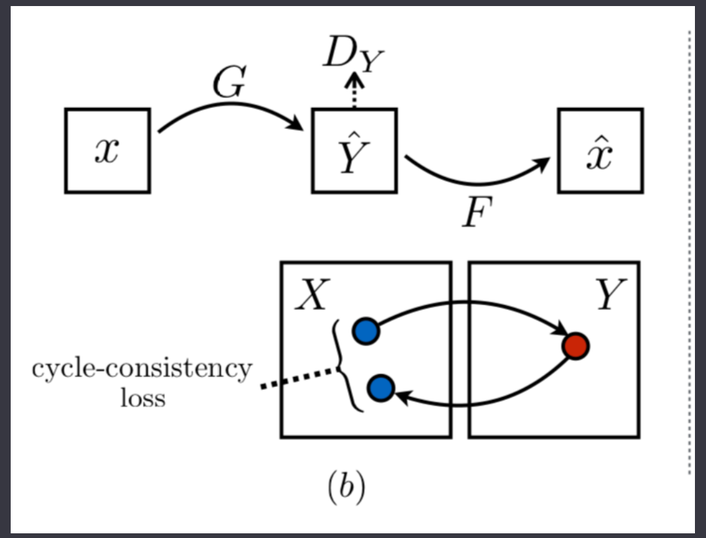
log( )

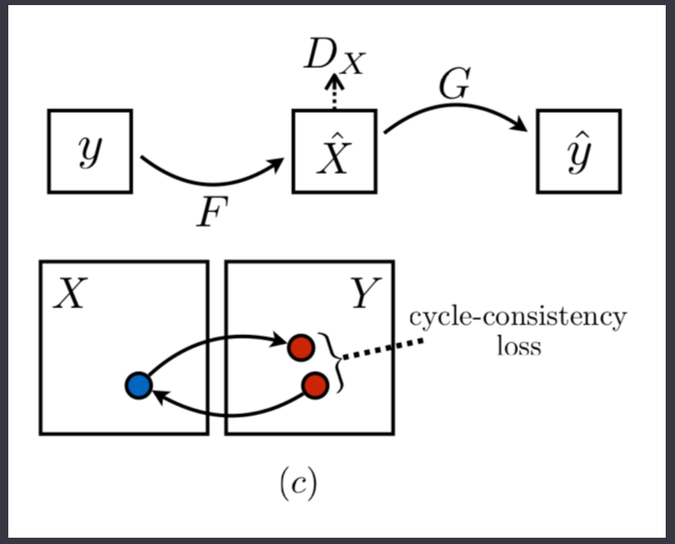
Cycle generative adversary network



Tweak on the architecture







$$\mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) = \mathbb{E}_{y \sim p_{\text{data}}(y)}[\log D_Y(y)] + \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log(1 - D_Y(G(x))], \quad (1)$$

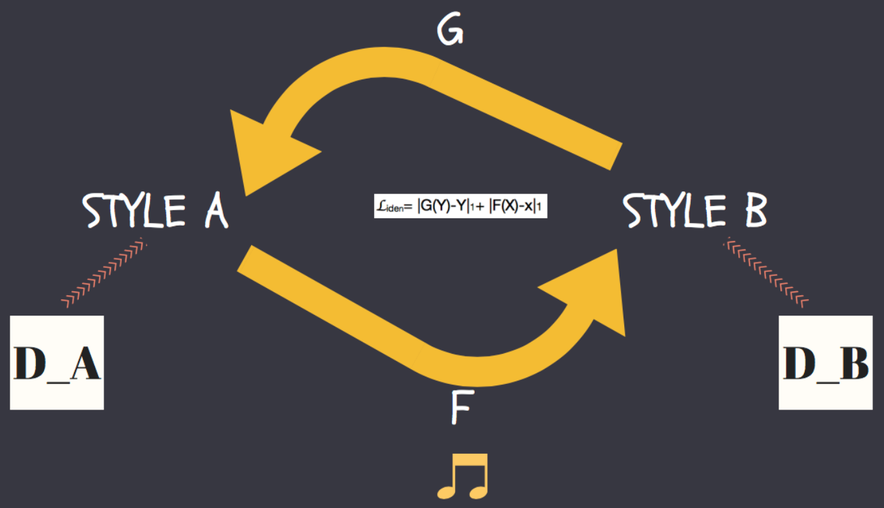
$$\mathcal{L}_{\text{cyc}}(G, F) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\|F(G(x)) - x\|_1] + \mathbb{E}_{y \sim p_{\text{data}}(y)}[\|G(F(y)) - y\|_1]. \quad (2)$$

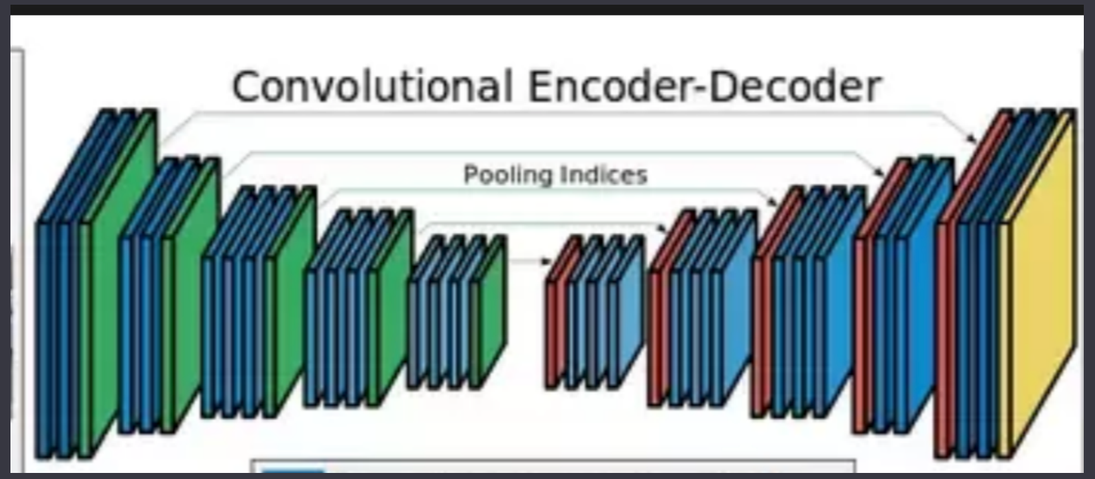
$$\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) + \mathcal{L}_{\text{GAN}}(F, D_X, Y, X) + \lambda \mathcal{L}_{\text{cyc}}(G, F), \quad (3)$$

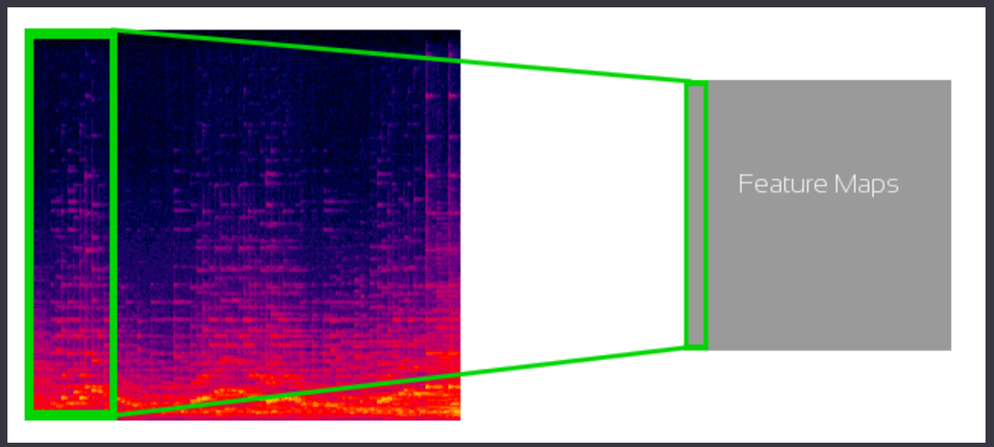
+

$$\mathcal{L}_{\text{iden}} = |G(Y) - Y|_1 + |F(X) - x|_1$$

Tweak on Generetor G and F







- 2 convolution layers (stride=2, width=3)+
- 9 Resnet layers (stride=1,width=3)+
- 2 transpose convolution layers +
- No pooling
- Instance normal each layer
- Relu as activation

Tweak on discriminators

- 5 convolution layers (stride=[2,2,2,2,1], width=4)+
- 1 dense layer
- No pooling
- Instance normal each layer
- Relu as activation