Jihong

Denoising Diffusion model variations

- DDPM → markov forward process
- DDIM → non-markov forward process

DDPM based diffusion model (dataset = CIFAR10)

- T = 300, epoch = 100
- 2min/1epoch → total = 2.5h (geforce RTX 4070) 오래 걸리는 건지 원래 이 정도 걸리는지 모르겠음
- loss : 0.58~0.6 (l1 loss) (한 50epoch 까지만 진행)
- 예시 코드에서는 loss 0.17~0.18수준 (dataset = Stanfordcars) / source가 삭제되서 CIFAR10로 대체

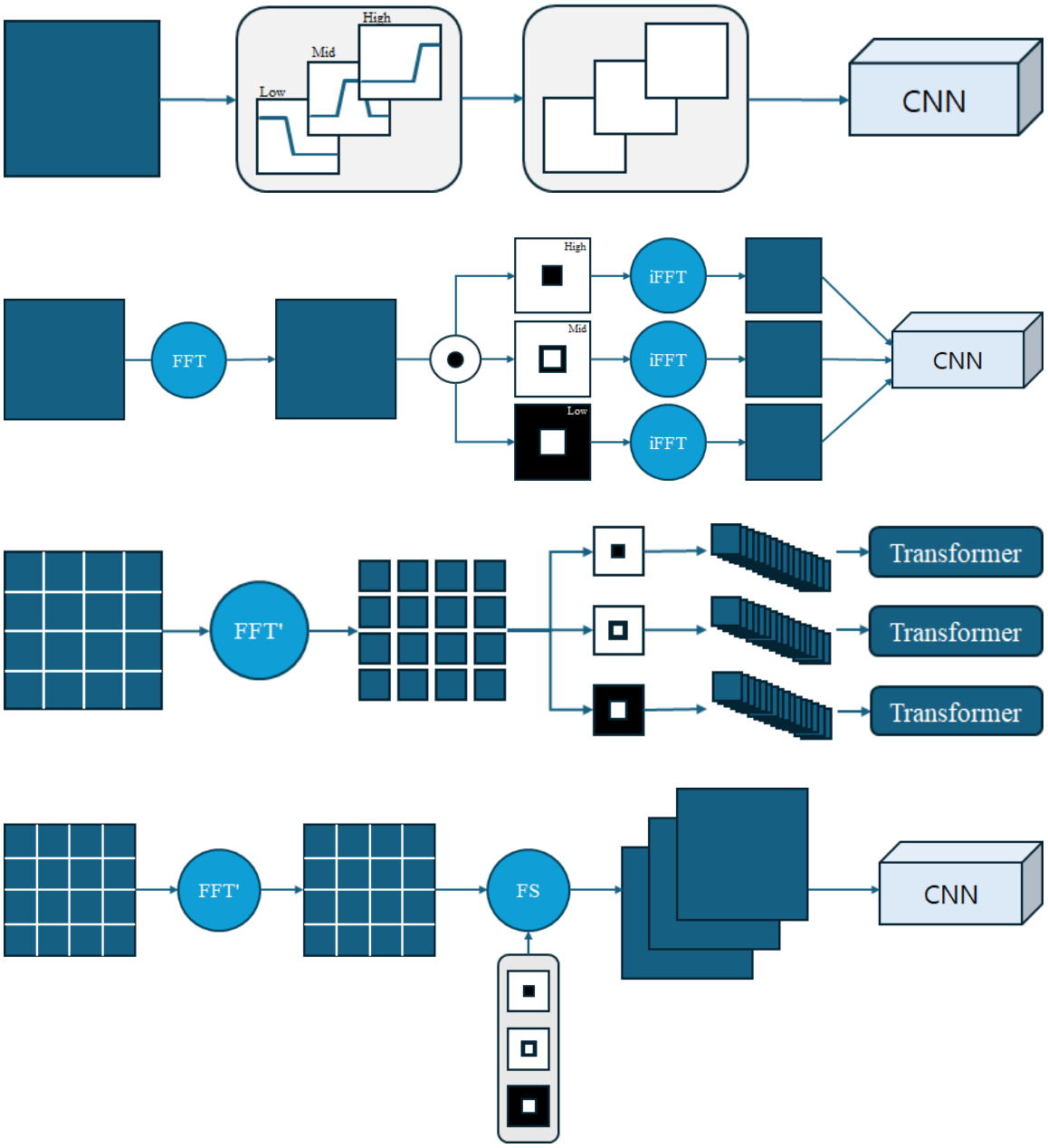
loss 계산할 때 noise/noised image vs model output

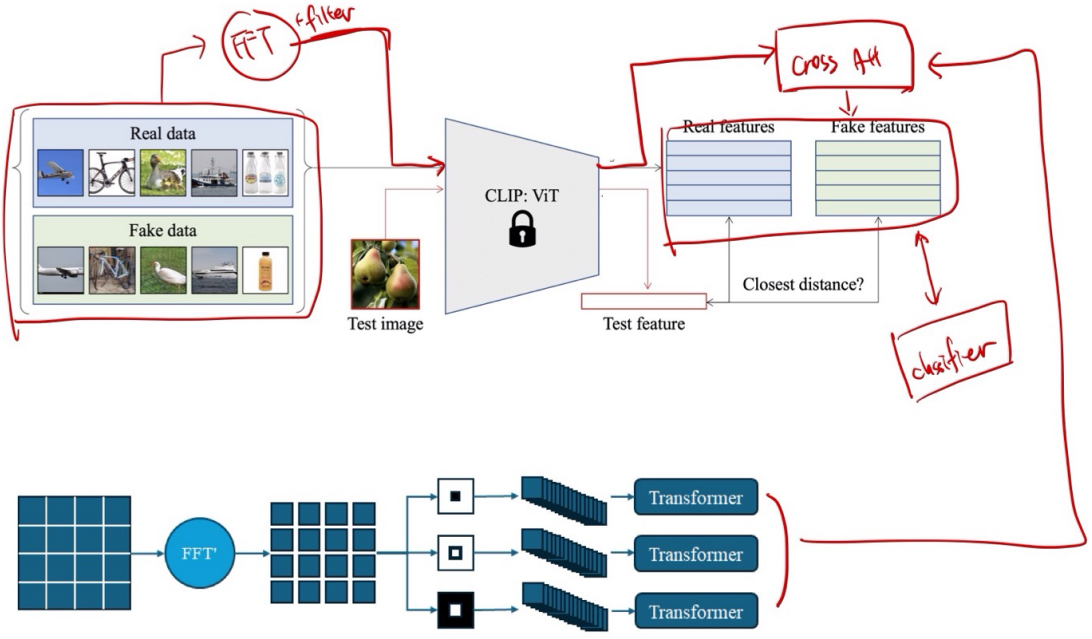
- noise랑 noised image 둘 중 뭐랑 비교하는 게 맞는지? → model output이 noised image를 prediction하는거니까 noised image랑 비교하는 게 맞는 것 같긴 한데..
- L1 loss는 noised image랑 비교할 때 더 낮음
- noise, noised image는 dim3(RGB), model output은 dim1 → 그대로 비교 or model output을 dim3으로 convolution하고 비교

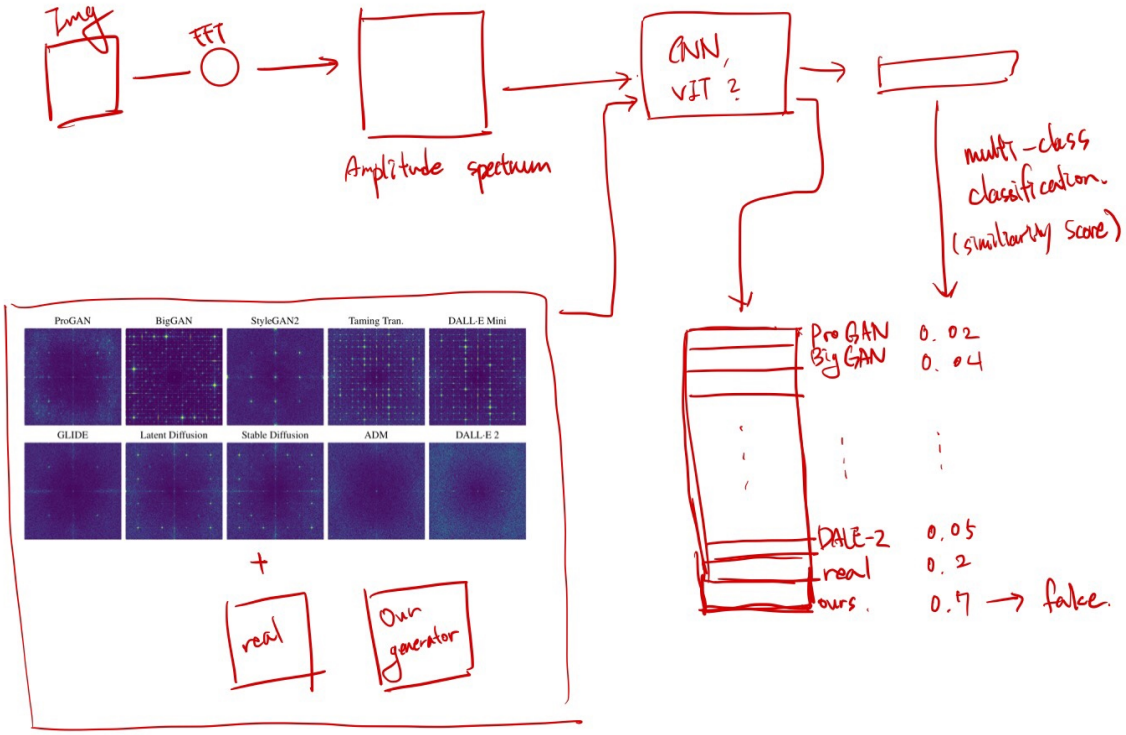
Generation은 fallow-up / Detection을 main task로

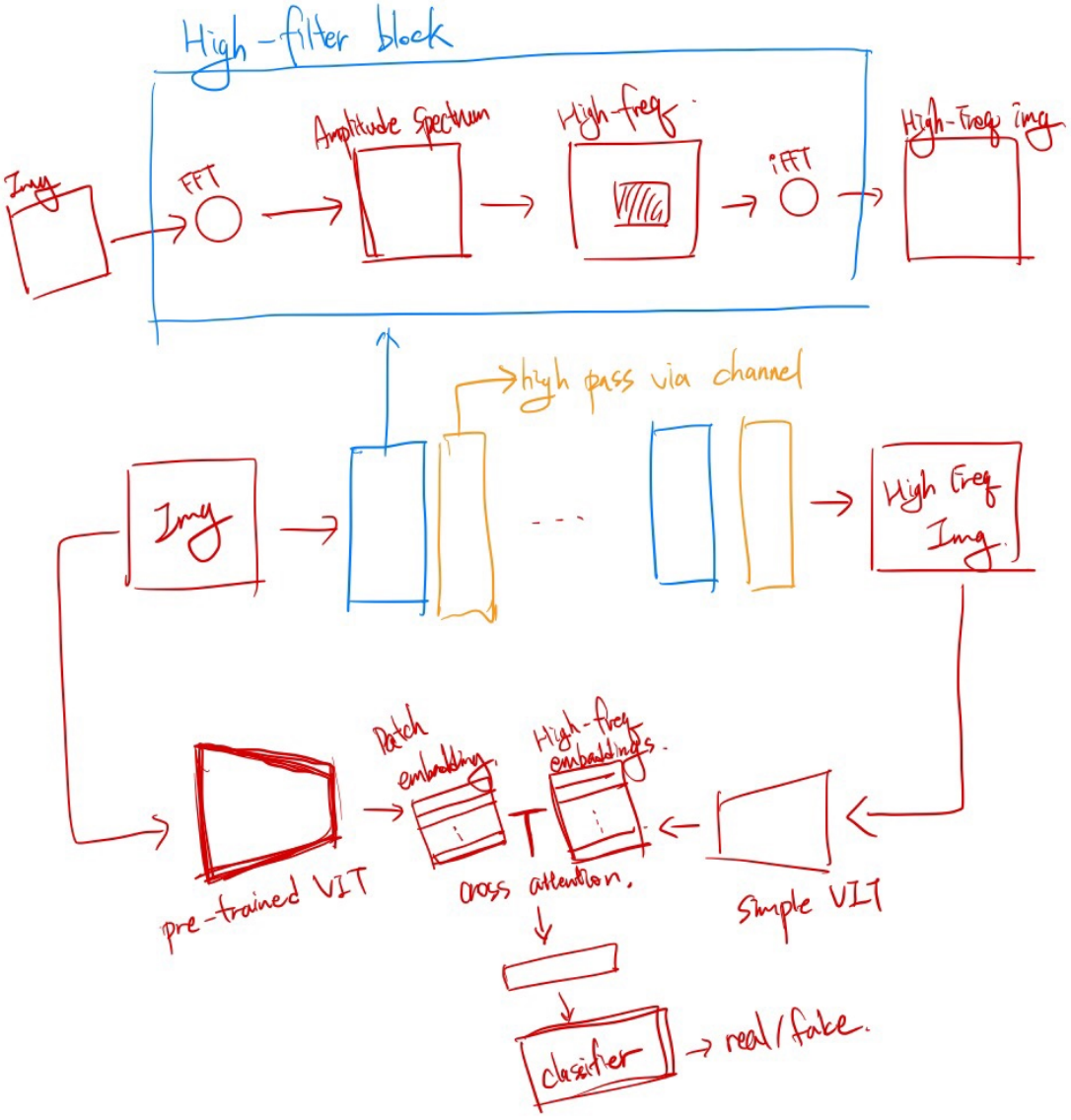
frequency-domain deepfake detections

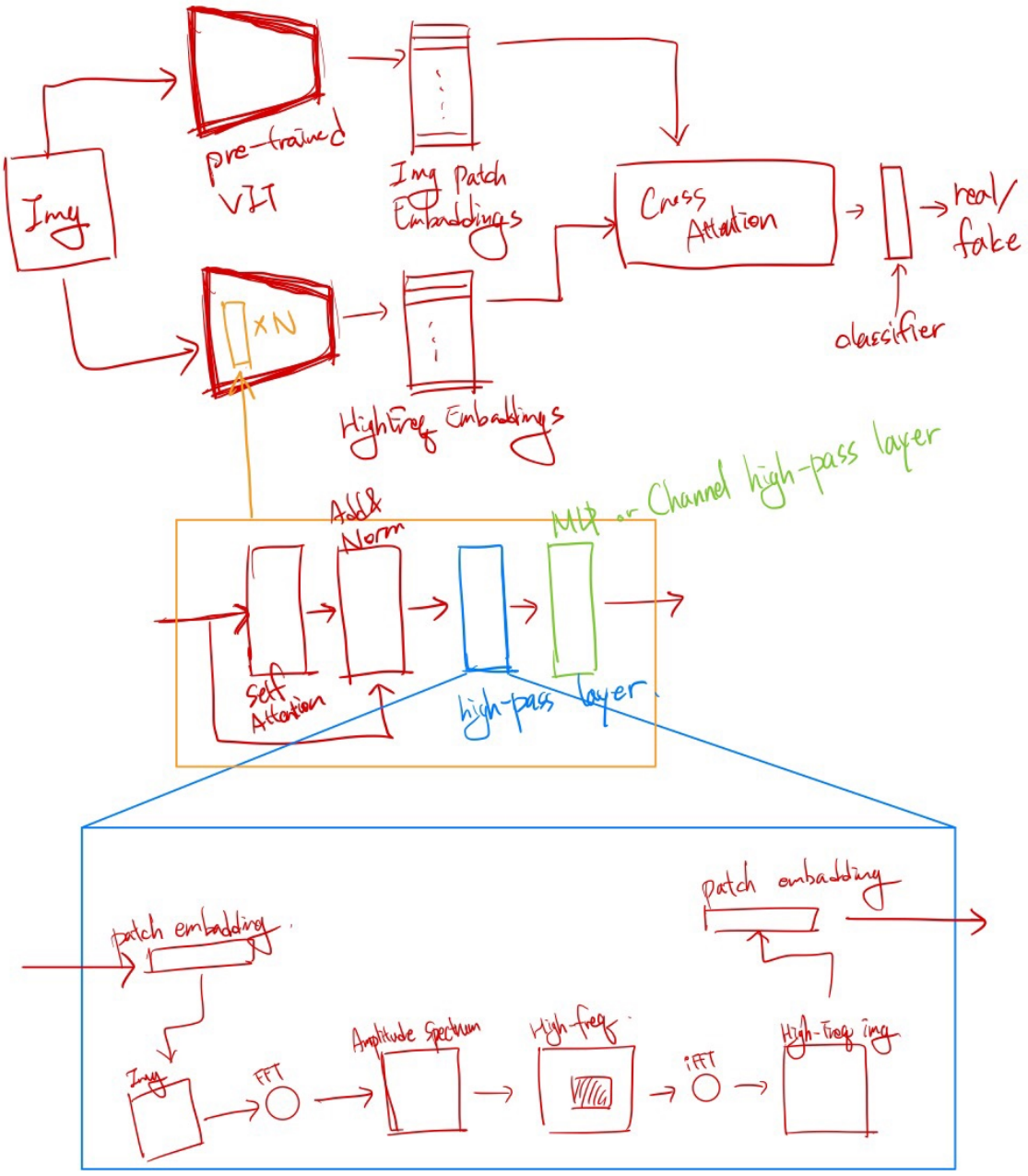
- FreqNet
- F3Net


High
Mid
Low
CNN
FFT
High
Mid
Low
iFFT
iFFT
iFFT
CNN
FFT'
Transformer
Transformer
Transformer
FFT'
FS
CNN


FFT
filter
Real data
Fake data
Test image
CLIP: ViT
Cross Att
Real features
Fake features
Closest distance?
Test feature
classifier
FFT'
Transformer
Transformer
Transformer


Img
FFT
Amplitude spectrum
CNN, ViT ?
multi-class classification.
(similarity score)
ProGAN
BigGAN
StyleGAN2
Taming Tran.
DALL·E Mini
GLIDE
Latent Diffusion
Stable Diffusion
ADM
DALL·E 2
+
real
Our generator
ProGAN 0.02
BigGAN 0.04
DALE-2 0.05
real 0.2
ours. 0.7 → fake.


High-filter block
Img
FFT
Amplitude Spectrum
High-freq
iFFT
High-freq img
high pass via channel
Img
High Freq Img
Patch embedding
High-freq embeddings
pre-trained VIT
cross attention
Simple VIT
classifier
real/fake



Frequency-based detection

- Frequency-level과 image-level analysis를 dual stream으로
- vit-base
- attention layer 처럼 frequency layer 만들어서 반복적인 학습이 가능하도록 하기