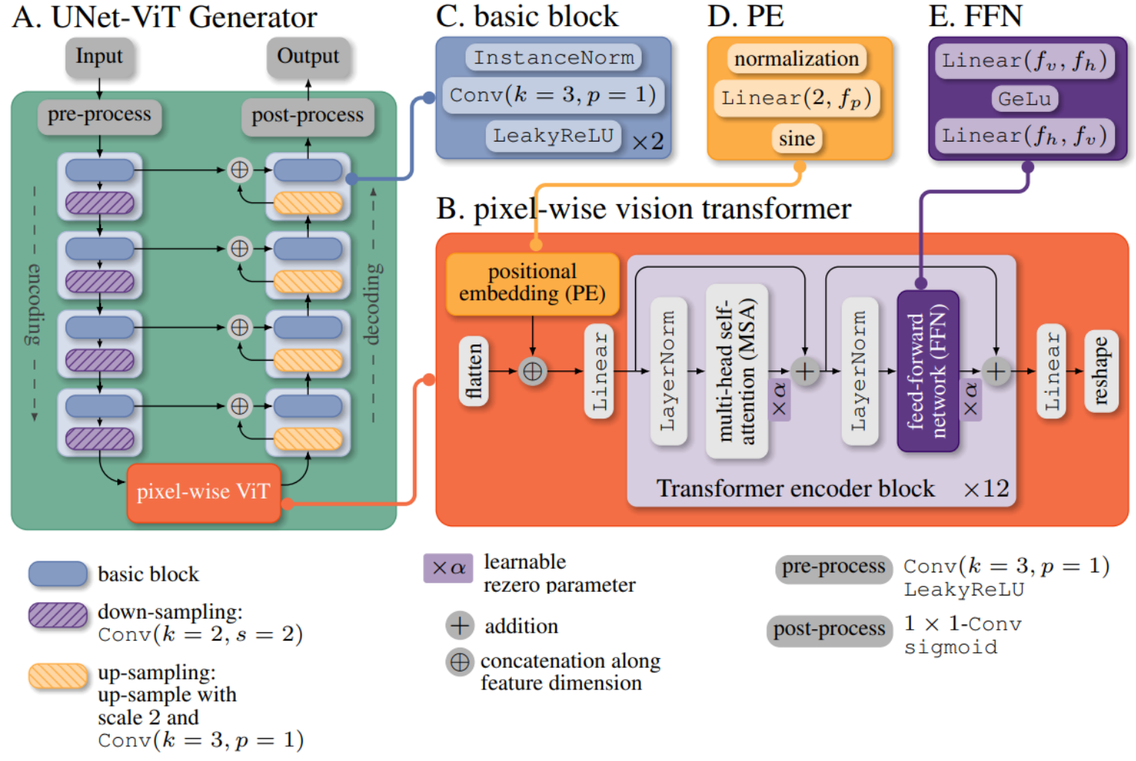
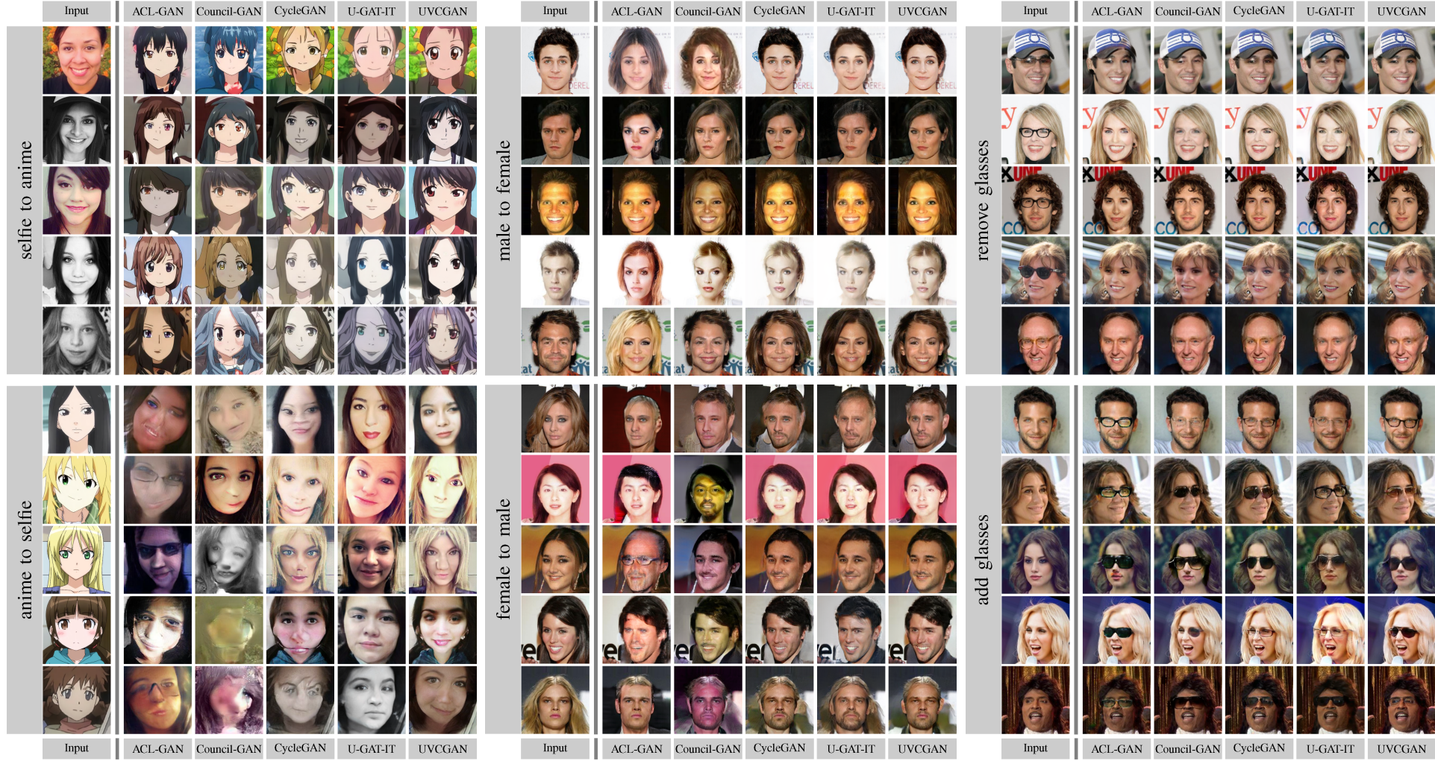

A. UNet-ViT Generator
Input
Output
pre-process
post-process
encoding
decoding
pixel-wise ViT
C. basic block
InstanceNorm
Conv(k = 3, p = 1)
LeakyReLU
×2
D. PE
normalization
Linear(2, f_p)
sine
E. FFN
Linear(f_v, f_h)
GeLu
Linear(f_h, f_v)
B. pixel-wise vision transformer
positional embedding (PE)
flatten
Linear
LayerNorm
multi-head self-attention (MSA)
×α
LayerNorm
feed-forward network (FFN)
×α
Linear
reshape
Transformer encoder block ×12
basic block
down-sampling: Conv(k = 2, s = 2)
up-sampling: up-sample with scale 2 and Conv(k = 3, p = 1)
×α learnable rezero parameter
addition
concatenation along feature dimension
pre-process Conv(k = 3, p = 1) LeakyReLU
post-process 1 × 1-Conv sigmoid


Input
ACL-GAN
Council-GAN
CycleGAN
U-GAT-IT
UVCGAN
selfie to anime
male to female
remove glasses
anime to selfie
female to male
add glasses