Supplementary for P2D: Plug and Play Discriminator for accelerating GAN frameworks

EfficientNet We extract 4 intermediate features from `tf_efficientnet_lite0` corresponding to spatial resolution $\{64, 32, 16, 8\}$. Classifier 1 in Table 1 corresponds to 64 resolution and so on.

CLIP For CLIP, we choose layers $\{4, 6, 9, 11\}$ and the 512 dimensional vector output. The vector is then passed into classifier 5 in Table 2.

Because CLIP is based on vision transformers, each intermediate feature has a token length of 50. For each layer, we first pass the [CLS] token into a 2-layer MLP with $LReLU$ to obtain $c \in \mathbb{R}^{192}$. We then resize the remaining 49 tokens into a $768 \times 7 \times 7$ resolution feature, and concatenate c to each spatial location, forming a $960 \times 7 \times 7$ feature. Because the features at all layers have the same dimension, we can use separate classifiers with the same specifications for each, see Table 2.

1. More results

We show more 256×256 resolution results on 4 common datasets in Figure 1.

FFHQ is a common benchmark for generative models containing 70k high quality 1024×1024 resolution images of faces.

Churches The LSUN Churches dataset contains around 1.2M images of outdoor church images at 256×256 resolution.

AFHQ is a dataset containing 512×512 resolution images of animal faces of cats, dogs, and wildlife animals. AFHQ contains 5000 images from each category, which we combine into a dataset of 15k.

Art Painting is a small dataset containing 1000 images of art paintings of resolution 512×512.

We resize all images to 256×256 resolution using `area` interpolation in PyTorch before training P2D.

EfficientNet			
Classifier 1	Classifier 2	Classifier 3	Classifier 4
Resblock(in=24, out=256, stride=2)	Resblock(in=40, out=512, stride=2)	Resblock(in=112, out=512, stride=2)	Resblock(in=320, out=512, stride=2)
Resblock(in=256, out=512, stride=2)	Resblock(in=512, out=512, stride=2)	Resblock(in=512, out=512, stride=2)	MinibatchStd()
Resblock(in=512, out=512, stride=2)	Resblock(in=512, out=512, stride=2)	MiniBatchStd()	FC(in=8192, out=512)
Resblock(in=512, out=512, stride=2)	MinibatchStd()	FC(in=8192, out=512)	LReLU
MinibatchStd()	FC(in=8192, out=512)	LReLU	FC(in=512, out=1)
FC(in=8192, out=512)	LReLU	FC(in=512, out=1)	
LReLU	FC(in=512, out=1)		
FC(in=512, out=1)			

Table 1. Architecture of classifiers for EfficientNet backbone.

CLIP	
Classifier1,2,3,4	Classifier 5
Resblock(in=960, out=768, stride=2)	FC(in=512, out=512)
Resblock(in=768, out=384, stride=1)	LReLU
Resblock(in=384, out=192, stride=1)	FC(in=512, out=512)
MiniBatchStd()	LReLU
FC(in=3072, out=512)	FC(in=512, out=512)
LReLU	LReLU
FC(in=512, out=1)	MiniBatchStd()
	FC(in=512, out=1)

Table 2. Architecture of classifiers for CLIP backbone.

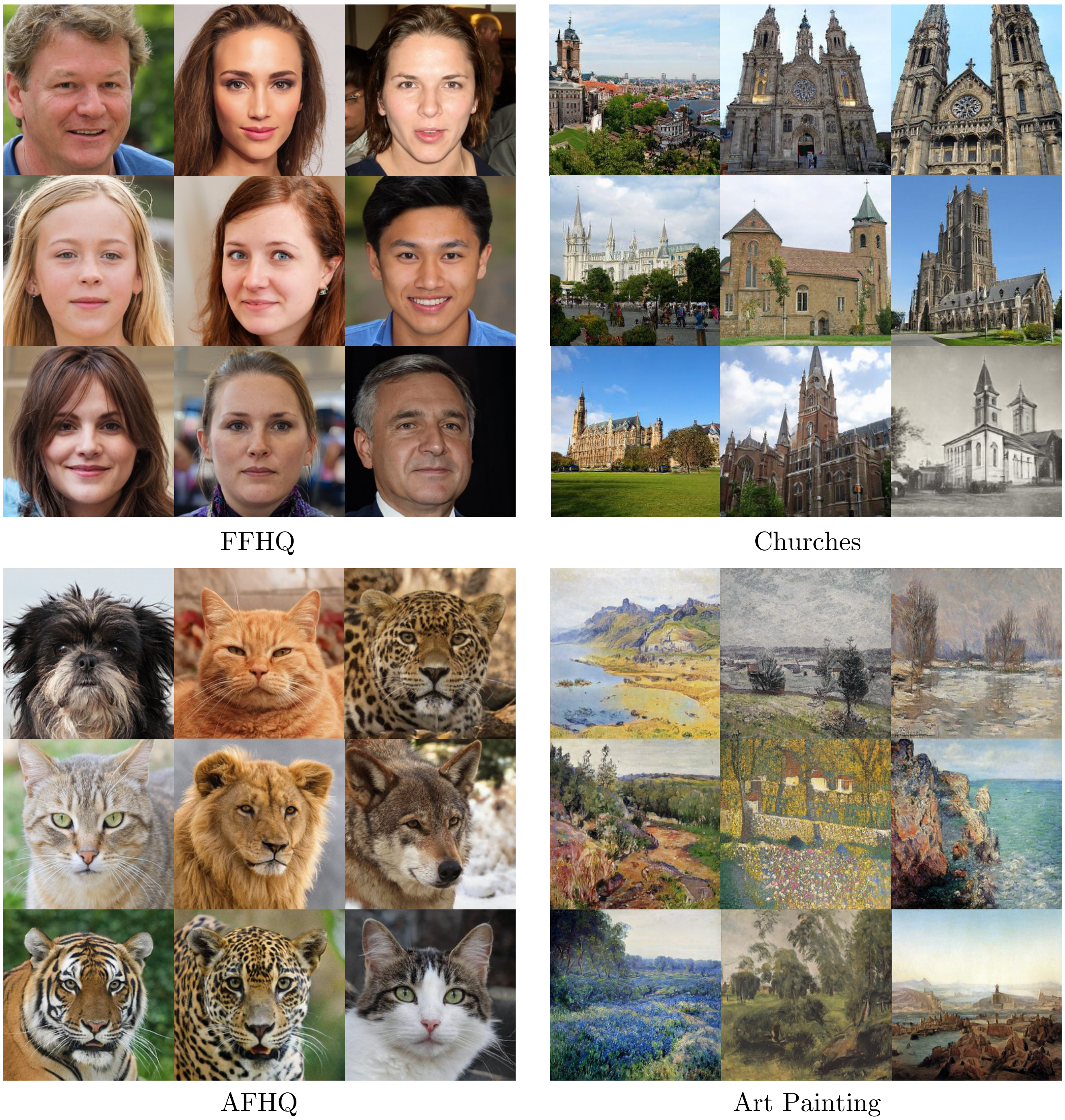

Figure 1. More random samples from P2D with StyleGAN2.