Fine-tuning StyleGAN-2 on Small Datasets

Ravish Rawal, Viren Bajaj

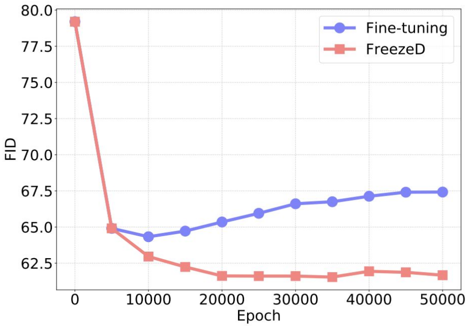
Executive Summary

1. Fine-tuning GANs is a good way to save **time** and **computation**, especially when our target dataset is **small** (1-10k images)
2. “GANs are notoriously tricky to train (Salimans et al., 2016)” - fine-tuning is also tricky because GANs **overfit very easily**, and there are a **multitude of pre-trained models** to start training from
3. We experiment with a new technique called freezeD (freezing the first D layers of the discriminator) to fine-tune StyleGAN2-ADA to a new dataset we created called BoredApes and **find the best freezeD configuration is when freezeD=10.**
4. StyleGAN2-ADA fine-tuned on the BoredApe dataset rather quickly - it could generate reasonable images in only 1K iterations as opposed to millions when training from scratch
5. We propose two new metrics: **inter-dataset similarity (IDS)** and **intra-dataset diversity (IDD)** to explore the relationship between the source and target dataset, and whether they correlate with convergence to a target KID
6. We find a positive correlation between IDS and freezeD, i.e., the more similar PT and FT datasets are to begin with, the more layers one can afford to freeze in fine tuning.
7. IDD results are inconclusive

Problem Motivation

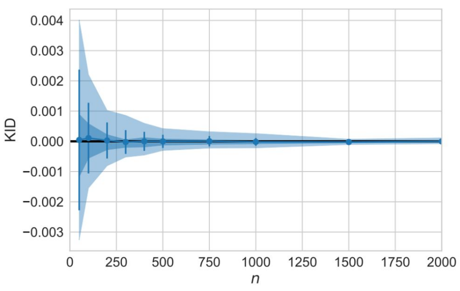
1. Training GANs from scratch requires a **large computational resources**, **data, and a lot of time** so fine-tuning pre-trained models is useful for data, resource, and time limited applications

2. **Fine-tuning** GANs is a good way to overcome this, but comes with it’s own problems:
    a. How do you prevent **overfitting** on a small dataset?
        i. Finetune both generator and discriminator? scale/shift normalization layers? GLO optimization? MineGAN? **FreezeD?**
        ii. FreezeD seems like a good baseline, but **how many layers should you freeze?**

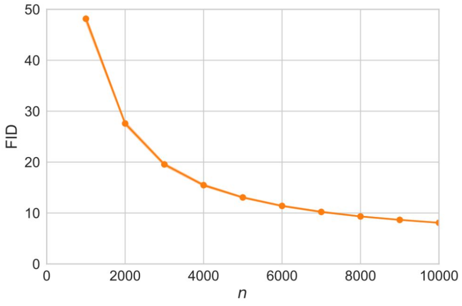


b. Given a target dataset, **which pre-trained model should you choose?**
   i. How **similar** is the source dataset to your target dataset?
   ii. How **diverse** was the original dataset?

c. What is the best metric to train on?
   i. Precision-Recall, Inception Score, Frechet Inception Distance, Kernel Inception Distance?



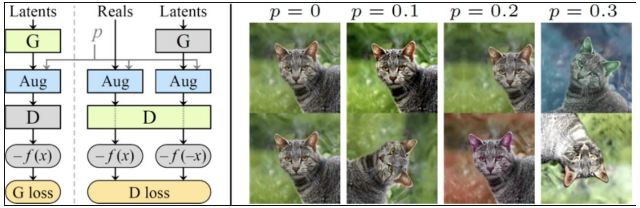
(a) KID estimates are unbiased, and standard deviations shrink quickly even for small n.



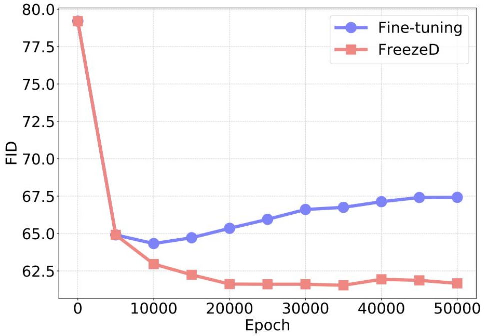
(b) FID estimates exhibit strong bias for n even up to 10 000. All standard deviations are less than 0.5.

Background Work

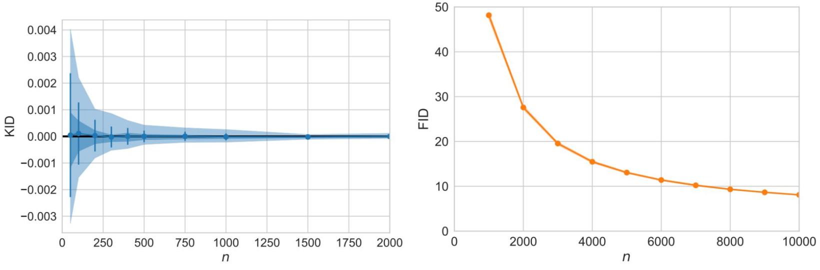
1. StyleGAN2: Training Generative Adversarial Networks with Limited Data (https://nvlabs-fi-cdn.nvidia.com/stylegan2-ada/ada-paper.pdf)



2. Freeze the Discriminator: a Simple Baseline for Fine-Tuning GANs (https://arxiv.org/abs/2002.10964)



3. Demystifying MMD GANs (https://arxiv.org/pdf/1801.01401.pdf)



(a) KID estimates are unbiased, and standard deviations shrink quickly even for small n.

(b) FID estimates exhibit strong bias for n even up to 10 000. All standard deviations are less than 0.5.

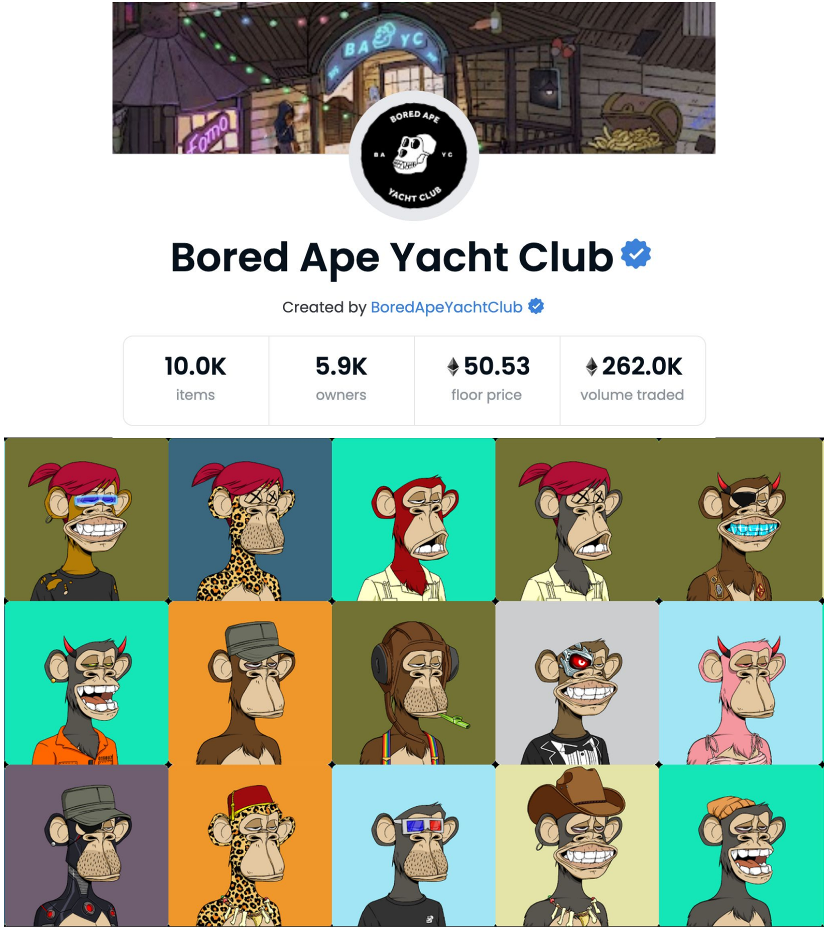
Technical Challenges

1. Curating a dataset:

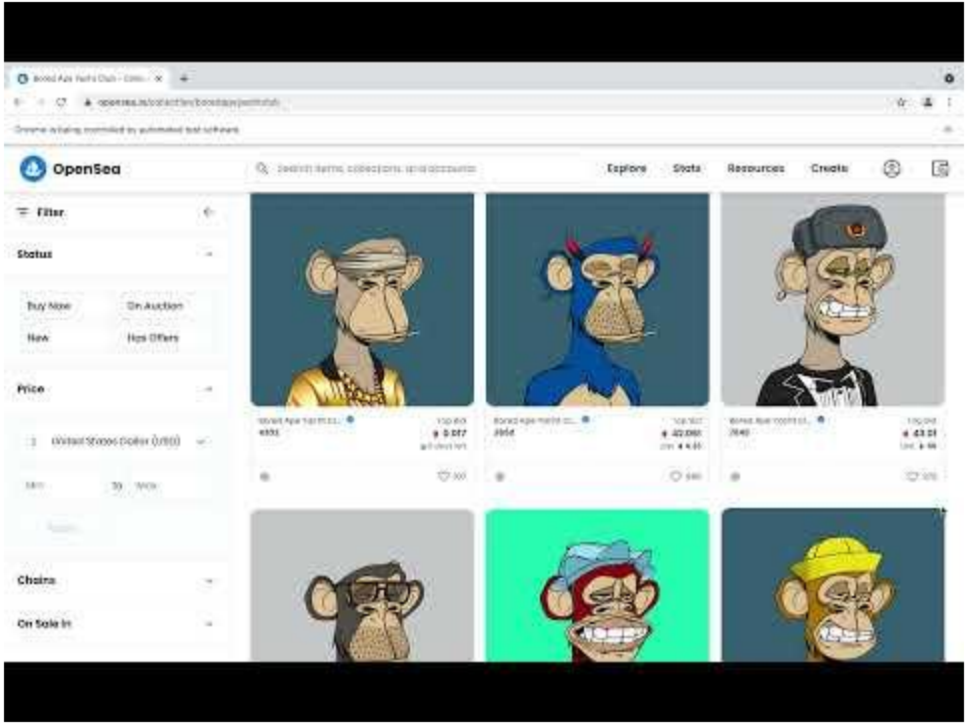
   Bored Ape Dataset

   a. Scraping
   b. Pre-processing

2. Transfer learning
   a. Choosing the right metric for a small dataset - FID or **KID**?
   b. Choosing the right pre-training dataset
   c. Preventing overfitting
      i. **freezeD** - freeze 10,11,12, or 13 layers of the StyleGAN2 Discriminator?



Scraping an infinite scroll webpage using Selenium to generate our dataset



1. Measuring effect of similarity between source and target datasets on KID

2. Measuring effect of diversity of the source dataset on KID

Picture: Two imaginary celebrities that were dreamed up by a random number generator.

Approach

Experiments:

Target dataset: boredape-256x256-small (1K imgs)

Source dataset: FFHQ, CELEBHQ, LSUN-DOG

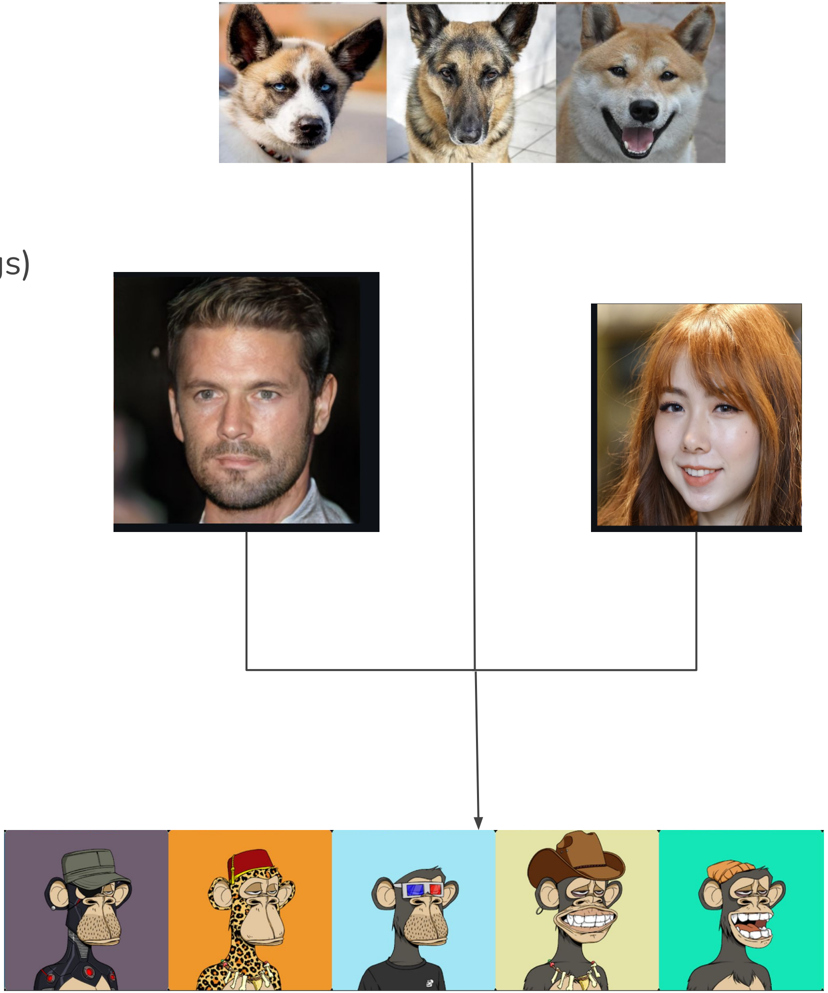
Model: StyleGAN2-ADA

FreezeD: 10,11,12,13

Evaluation Metric: KID

Similarity Metric: Inter-Dataset Similarity (IDS)

Diversity Metric: Intra-Dataset Diversity (IDD)



1. Plot KID vs FreezeD for StyleGAN2 pre-trained on [FFHQ, CELEBHQ, LSUNCAT] by training on 1000 real images from BoredApe

2. Calculate similarity between BoredApe and [FFHQ, CELEBHQ, LSUNCAT]

3. Calculate Diversity of BoredApe and [FFHQ, CELEBHQ, LSUNCAT]

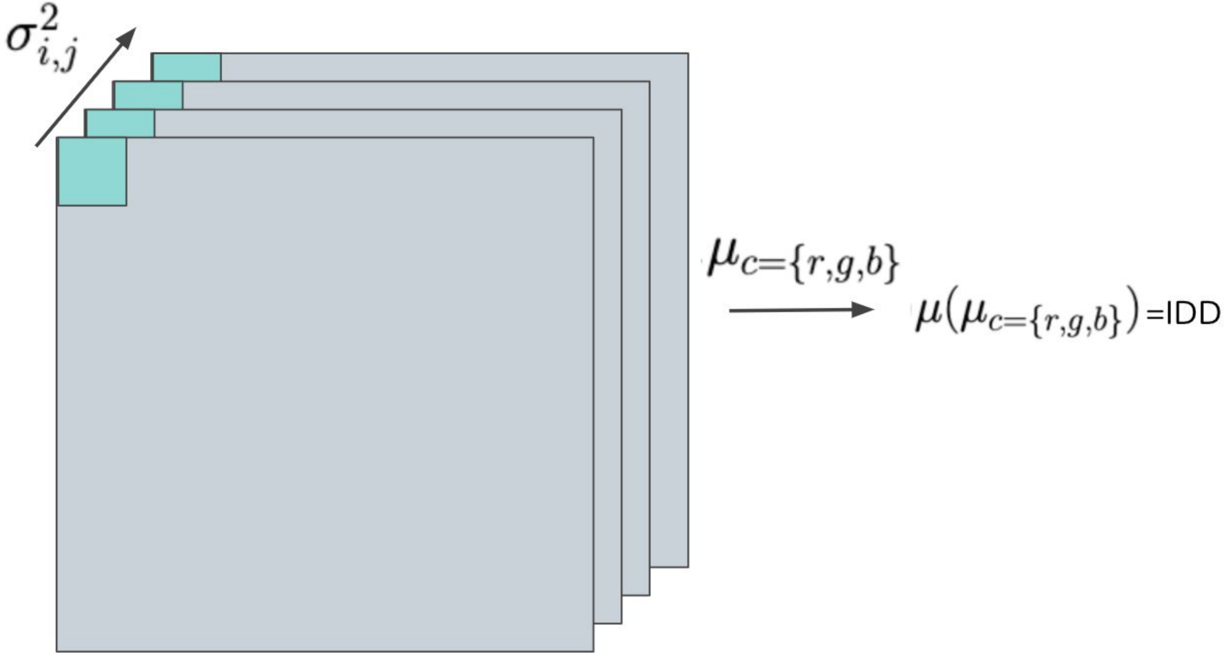
“High IDD and high IDS imply better quality of transfer learning (lower KID)”

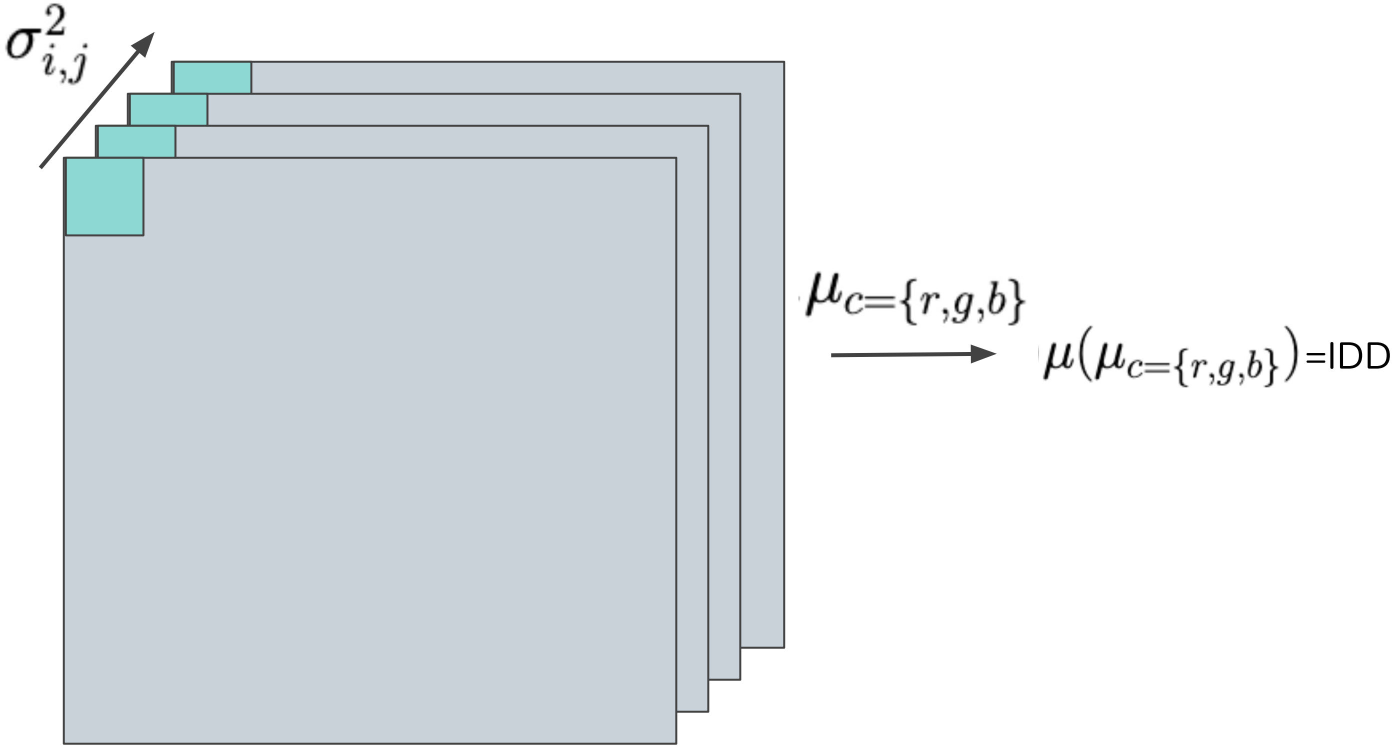
In order to assess which pre-training (PT) dataset would work best with our fine-tuning (FT) data, and to provide the community with a guideline for assessing best fit beforehand in the future, we developed two metrics: Intra-Dataset Diversity (IDD) and Inter-Dataset Similarity (IDS).

- IDD: Past work has mentioned that diversity of the PT dataset improves KID performance of StyleGAN, but this is a visual heuristic. We developed the IDD metric to quantify diversity within a dataset by analyzing the pixel-wise distribution of RGB values along the depth of the dataset, and flattening these to an average. This encodes positional and pixel value data

- IDS: Similarity between PT & FT datasets has long been known to improve performance. We used inverse KID to measure similarity between our datasets.

Calculating IDD





$\sigma^2_{i,j}$
$\mu_{c=\{r,g,b\}}$
$\mu(\mu_{c=\{r,g,b\}})$=IDD

StyleGAN2-ADA Architecture

Style Based Generator

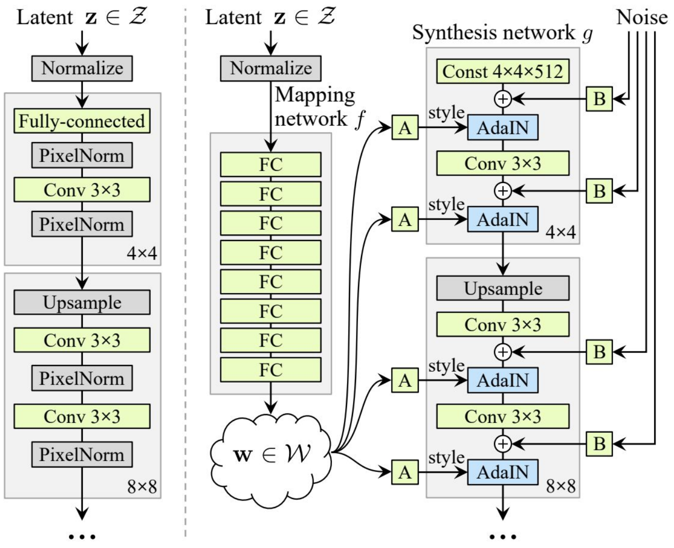
1. Traditional generator - feeds the latent code though the input layer only,
   Style Based - maps the input to an intermediate latent space W, which then controls the generator through adaptive instance normalization (AdaIN) at each convolution layer.

2. “A” = learned affine transform, and
   “B” = learned per-channel scaling factors to the noise input

3. Mapping network f - 8 layers
   Synthesis network - g consists of 18 layers two for each resolution ($4^2 - 1024^2$).

4. The output of the last layer is converted to RGB using a separate 1 × 1 convolution,

5. Trainable Parameters in style based = 26.2M Trainable parameters in traditional 23.1M



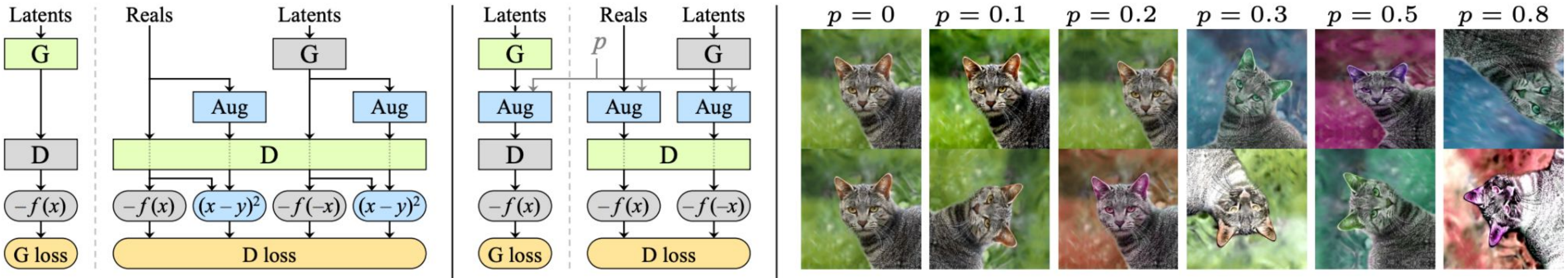
(a) Traditional (b) Style-based generator

Discriminator with Adaptive Discriminator Augmentation:

1. (a) Balanced consistency regularization (bCR) (b) stochastic discriminator augmentations

2. (b) StyleGAN2-ADA uses the non-saturating logistic loss f(x) = log (sigmoid(x))

3. (c) The effect of a diverse set of augmentations to every image that the discriminator sees, controlled by an augmentation probability p.



(a) bCR (previous work) (b) Ours (c) Effect of augmentation probability p

Implementation Details

Implementation Details

Fine-tuning StyleGAN2-ADA example:

```
python train.py
--outdir=./training-runs
--data=boredape-small-256x256.z
ip --gpus=1 --resume=ffhq256
--kimg=1000 --metrics=none
--batch=32 --gamma=2
--freezed=13 --snap=10
```

GPUs used = A100, V100

Time to train for 1000 iterations: 3 hr (A100), 4.34 hr (V100)

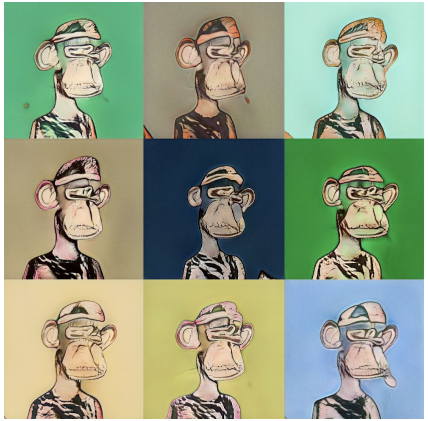

```
Training options:
{
  "num_gpus": 1,
  "image_snapshot_ticks": 10,
  "network_snapshot_ticks": 10,
  "metrics": [],
  "random_seed": 0,
  "training_set_kwargs": {
    "class_name": "training.dataset.ImageFolderDataset",
    "path": "boredape-small-256x256.zip",
    "use_labels": false,
    "max_size": 1000,
    "xflip": false,
    "resolution": 256
  },
  "data_loader_kwargs": {
    "pin_memory": true,
    "num_workers": 3,
    "prefetch_factor": 2
  },
  "G_kwargs": {
    "class_name": "training.networks.Generator",
    "z_dim": 512,
    "w_dim": 512,
    "mapping_kwargs": {
      "num_layers": 2
    },
    "synthesis_kwargs": {
      "channel_base": 16384,
      "channel_max": 512,
      "num_fp16_res": 4,
      "conv_clamp": 256
    }
  },
  "D_kwargs": {
    "class_name": "training.networks.Discriminator",
    "block_kwargs": {
      "freeze_layers": 13
    },
    "mapping_kwargs": {},
    "epilogue_kwargs": {
      "mbstd_group_size": 4
    },
    "channel_base": 16384,
    "channel_max": 512,
    "num_fp16_res": 4,
    "conv_clamp": 256
```

```
  "G_opt_kwargs": {
    "class_name": "torch.optim.Adam",
    "lr": 0.0025,
    "betas": [
      0,
      0.99
    ],
    "eps": 1e-08
  },
  "D_opt_kwargs": {
    "class_name": "torch.optim.Adam",
    "lr": 0.0025,
    "betas": [
      0,
      0.99
    ],
    "eps": 1e-08
  },
  "loss_kwargs": {
    "class_name": "training.loss.StyleGAN2Loss",
    "r1_gamma": 2.0
  },
  "total_kimg": 1000,
  "batch_size": 32,
  "batch_gpu": 32,
  "ema_kimg": 5.0,
  "ema_rampup": null,
  "ada_target": 0.6,
  "augment_kwargs": {
    "class_name": "training.augment.AugmentPipe",
    "xflip": 1,
    "rotate90": 1,
    "xint": 1,
    "scale": 1,
    "rotate": 1,
    "aniso": 1,
    "xfrac": 1,
    "brightness": 1,
    "contrast": 1,
    "lumaflip": 1,
    "hue": 1,
    "saturation": 1
  },
```

Demo

StyleGAN Pretrained on FFHQ Learns Fast

Initial fakes



20 Trained Images

1000 Trained Images

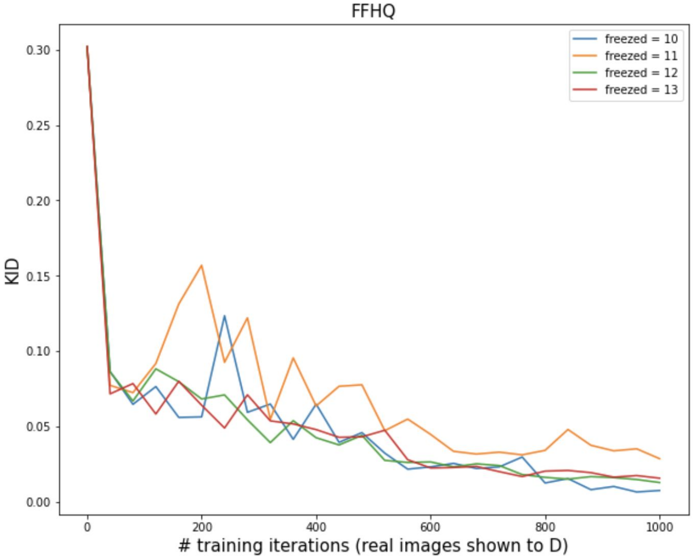
Reals

Experimental Evaluation

IDD & IDS Scores Relative to Bored Apes

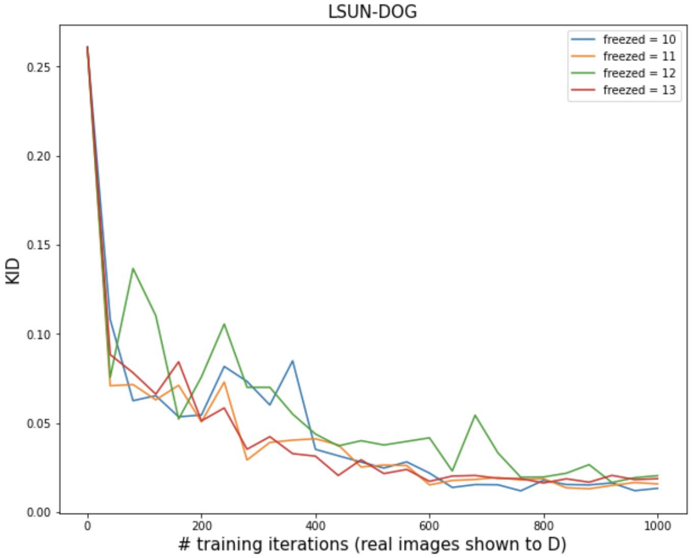
Dataset	IDD (Higher = Better)	IDS (Higher = Better)
FFHQ	46.44	3.85
CELEBAHQ	59.08	3.33
LSUN Dogs	64.91	5.26
Bored Apes	62.80	Infinity
Random Noise	73.57	1.47

1. **Results are more varied when datasets are less similar implying that we can afford to freeze more D layers when the IDS score is high.**
2. **FreezeD=10 is the best for Fine-Tuning on the BoredApe Dataset from FFHQ and LSUN-DOG**



0.0076
0.0287
0.0129
0.0158

10 < 12 < 13 < 11

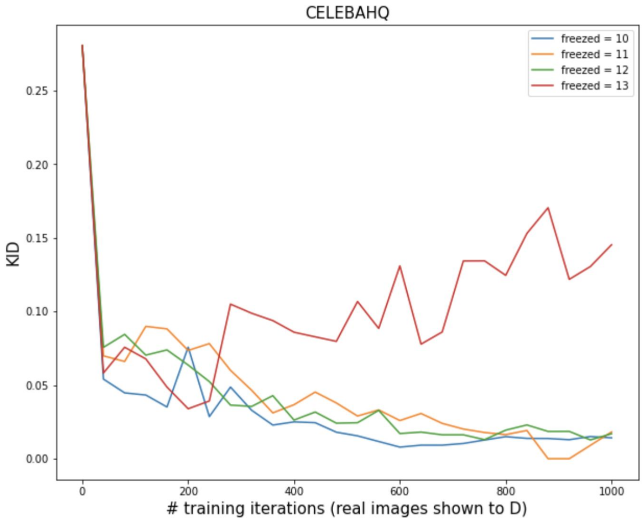


0.0132
0.0158
0.0204
0.0188

10 < 11 < 13 < 12

Pretrained on: FFHQ LSUN-DOG

1. **Results are more varied when datasets are less similar**
2. **FreezeD=10 is the best for CELEBAHQ as well, 13 diverges even though it was recommended**
3. **IDD results are inconclusive**



0.0142
0.0182
0.0169
0.1453

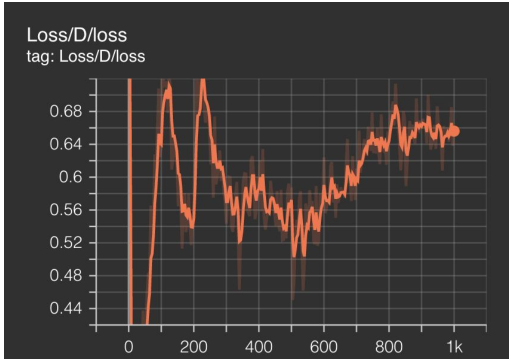
10 < 12 < 11 < 13

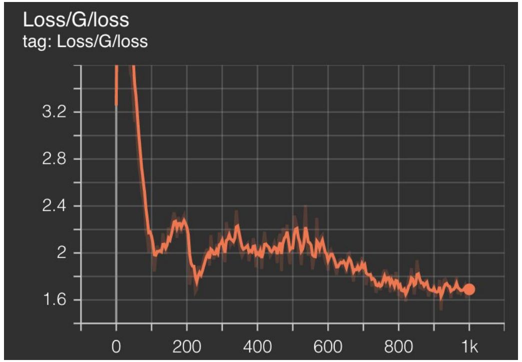
Pretrained on: CELEBAHQ

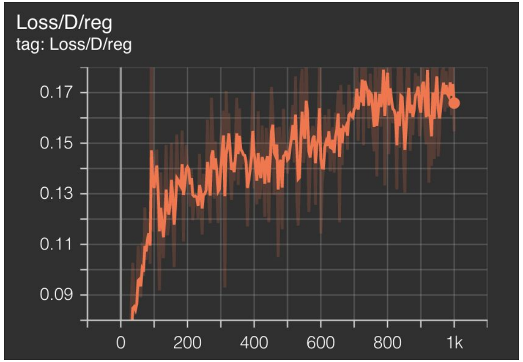
Conclusion

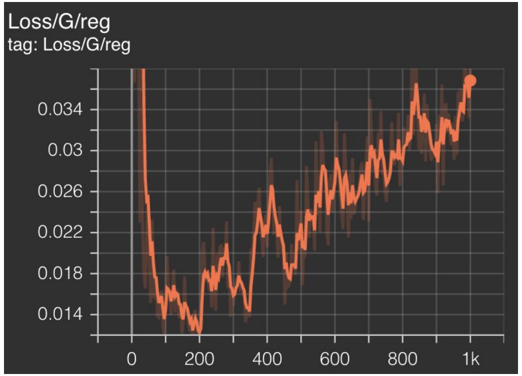
1. StyleGAN2-ADA fine-tuned on the BoredApe dataset rather quickly - it could generate reasonable images in only 1K iterations as opposed to millions when training from scratch
2. FreezeD = 10 works the best in all our finetuning experiments
3. We introduce new metrics to compare the diversity of the source dataset and the similarity between the source and target dataset: IDD and IDS
4. We find that KID results are more varied at convergence when datasets are less similar, implying that we can afford to freeze more D layers when the IDS score is high.
5. IDD results are inconclusive

D/G Loss and Regularization

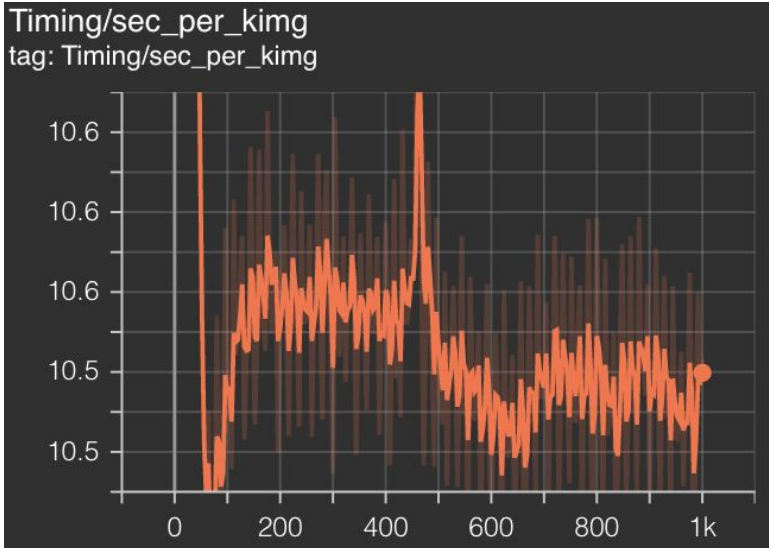


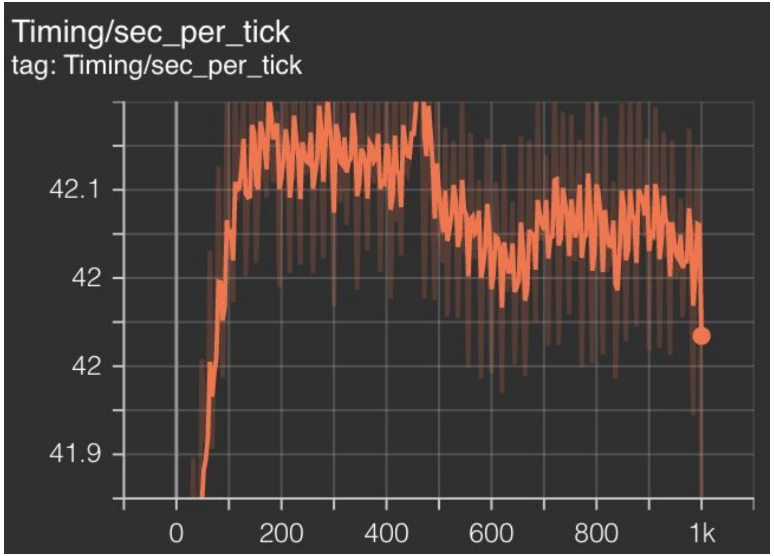




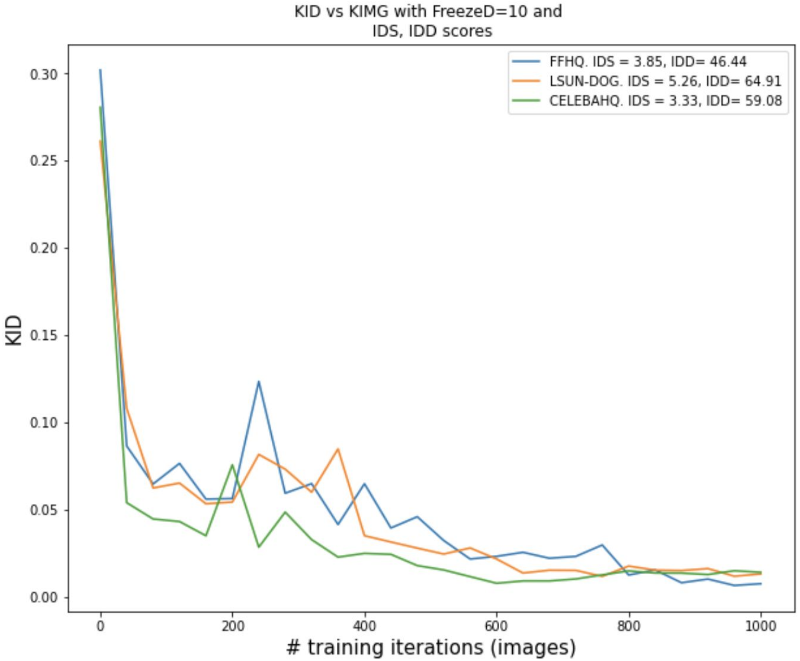


Timing for FFHQ Freeze10 on A100





FreezeD=10 performs well for all datasets



0.0076
0.0132
0.0142

FFHQ <
LSUN-DOG <
CELEBAHQ

FreezeD=13 Diverges even though it was recommended