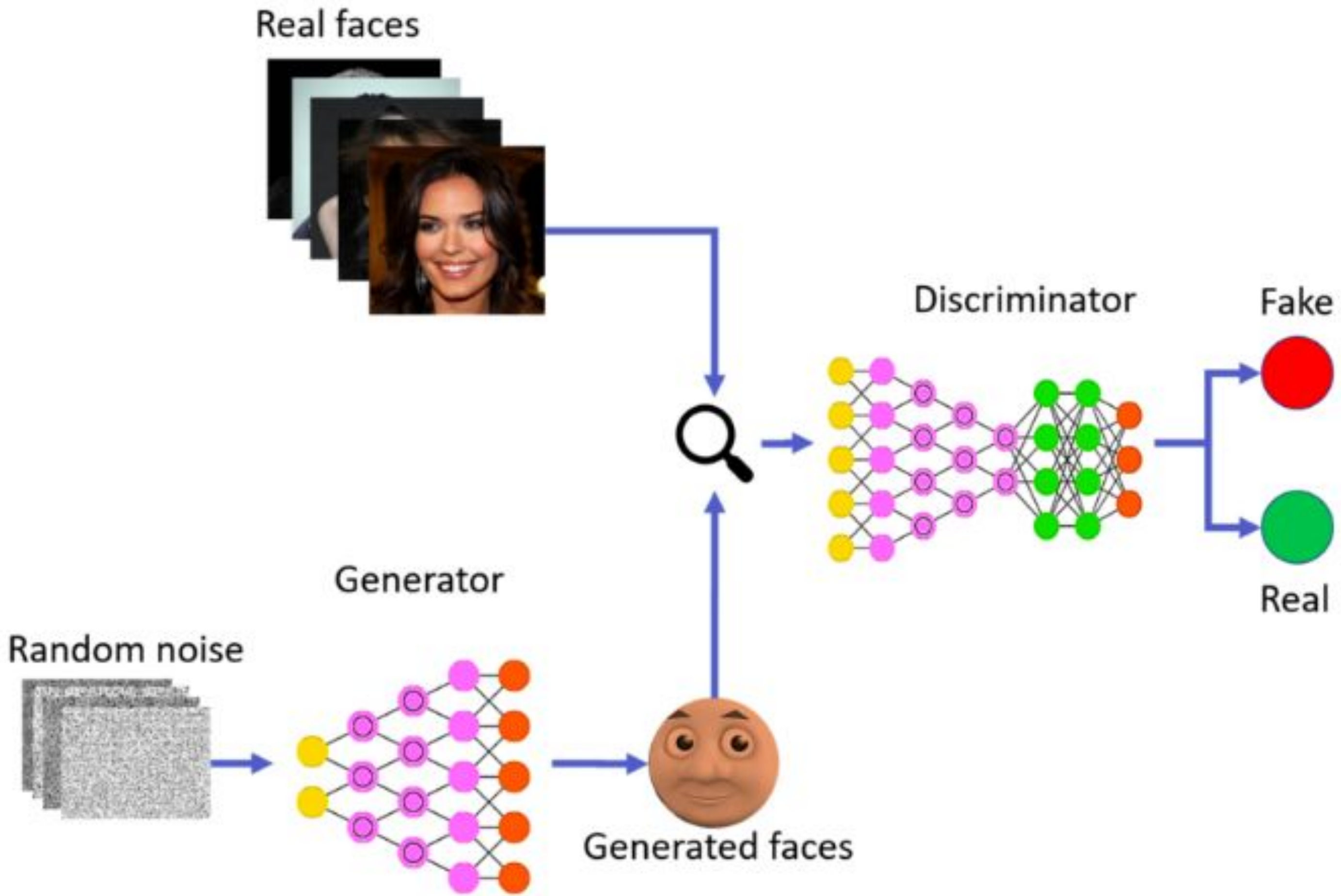
PokéGAN: Generating Pokémon Sprites Using a Generative Adversarial Network

By Braeden Alonge and Lucas Summers

Motivation

- Pokémon: 1,000+ unique creatures with a distinct artistic style
  - So…can we use AI to make new ones just as good?
- **Challenge:** Generate novel pixel sprites indistinguishable from real ones
- **Solution:** GANs!!! (specifically Deep Convolutional GANs)
- **Goal:** 64x64 RGB images evaluated by FID, IS, and Diversity Score


Real faces
Discriminator
Fake
Real
Generator
Random noise
Generated faces

1,000 Pokémon Dataset (Kaggle)

- 26,000+ images of 1,000 different Pokémon
- **Split:** 90% train / 5% validation / 5% test
- **Preprocessing:**
  - Resize from 128x128 ⟶ 64x64
  - RGBA ⟶ RGB
  - Normalize to [-1, 1] range
- **Data augmentation:**
  - Horizontal flips (50% prob)
  - Vertical flips (10% prob)
  - Random rotations (5 deg)

Baseline Recap: Vanilla DCGAN

- **Generator** (~3.5M parameters)
  - 100-dim noise ⟶ 4×4×512 ⟶ progressive upsampling ⟶ 64×64×3 image
  - 4 layers of TransposeConv2d + BatchNorm + ReLU
  - Tanh output: [-1, 1]
- **Discriminator** (~2.5M parameters)
  - 64×64×3 image ⟶ progressive downsampling ⟶ probability score (real or fake)
  - 4 layers of Conv2d + BatchNorm + LeakyReLU(0.2)
  - Sigmoid output: [0, 1]
- **Training**: BCE loss, Adam (lr=0.0002), batch size 64, 200 epochs

Model Improvements

- **Self-Attention (SAGAN-style):** Captures long-range dependencies in pixels
- **Spectral Normalization:** Stabilize discriminator training
- **Label Smoothing (0.2):** Prevent discriminator overconfidence
- **Asymmetric Learning Rates (lr_D = 4x lr_G):** D learns faster and gives better info to G
- **Label Flipping (5% prob):** Add noise to prevent overfitting
- **Data Augmentation:** Randomized flips + rotations
- **Early Stopping:** Patience = 20 epochs, min_delta = 1.0 FID

Self-Attention Mechanism

- **Problem:** Early conv layers have limited context
- **Solution:** Every pixel attends to every other pixel
- **Implementation:**
  - Query, Key, Value projections (1x1 convs)
  - Output = gamma*(V*Attention)+x
  - Gamma starts at 0 (gradual incorporation of attention)
- Placed at 32x32 resolution in both G and D
- Hopefully will coordinate distant features (body parts, colors, etc.)

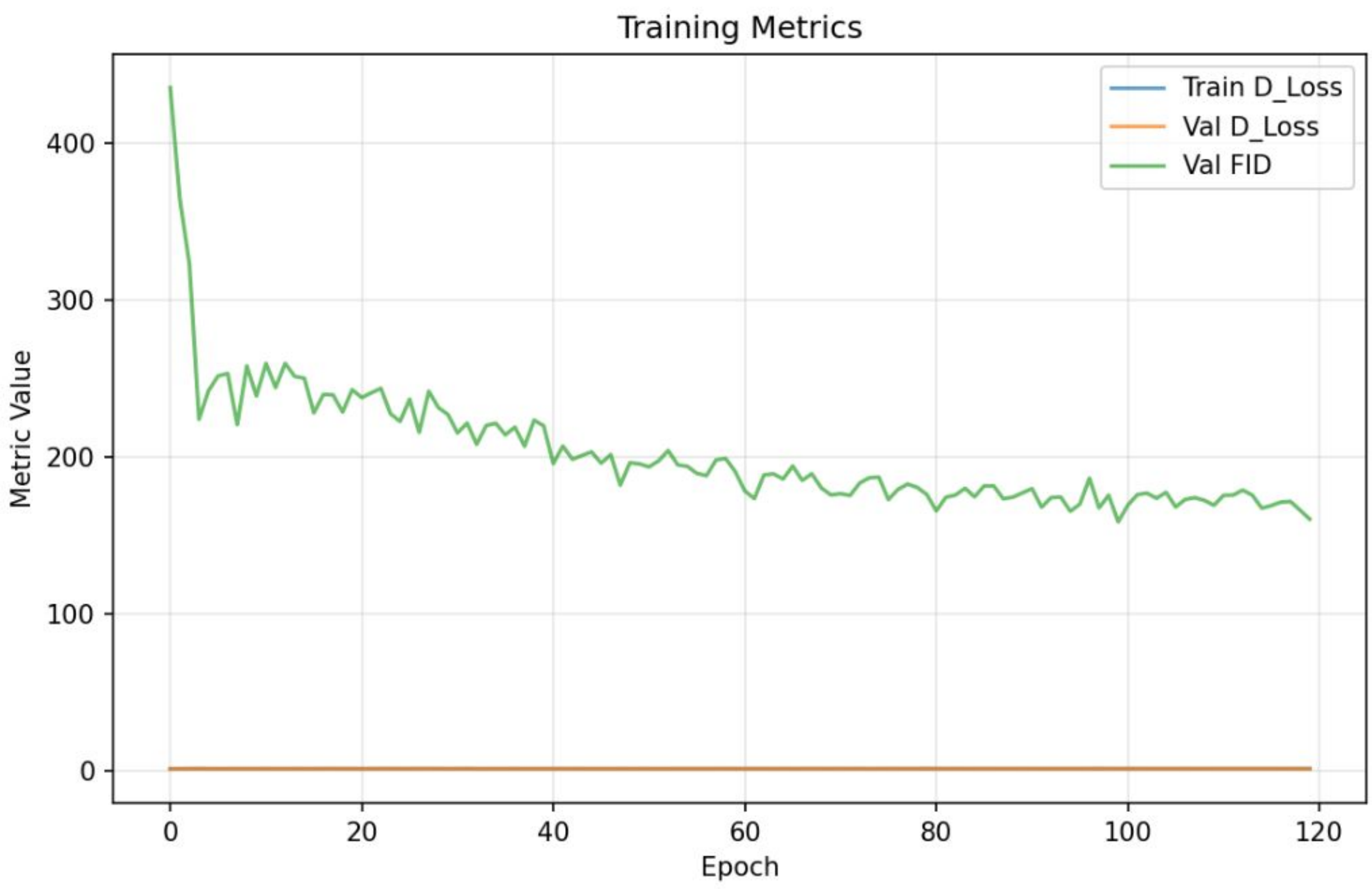
Results

Metric	Baseline Model	Final Model	Change
FID	218.59	79.92	-138.67
Inception Score	2.10±0.06	2.95 ± 0.12	+0.85
Diversity Score	52.84	59.81	+6.97
Discriminator Acc.	22.7%	62.95%	+40.25%
Epoch	200	99 (early stopping)	-101

Results

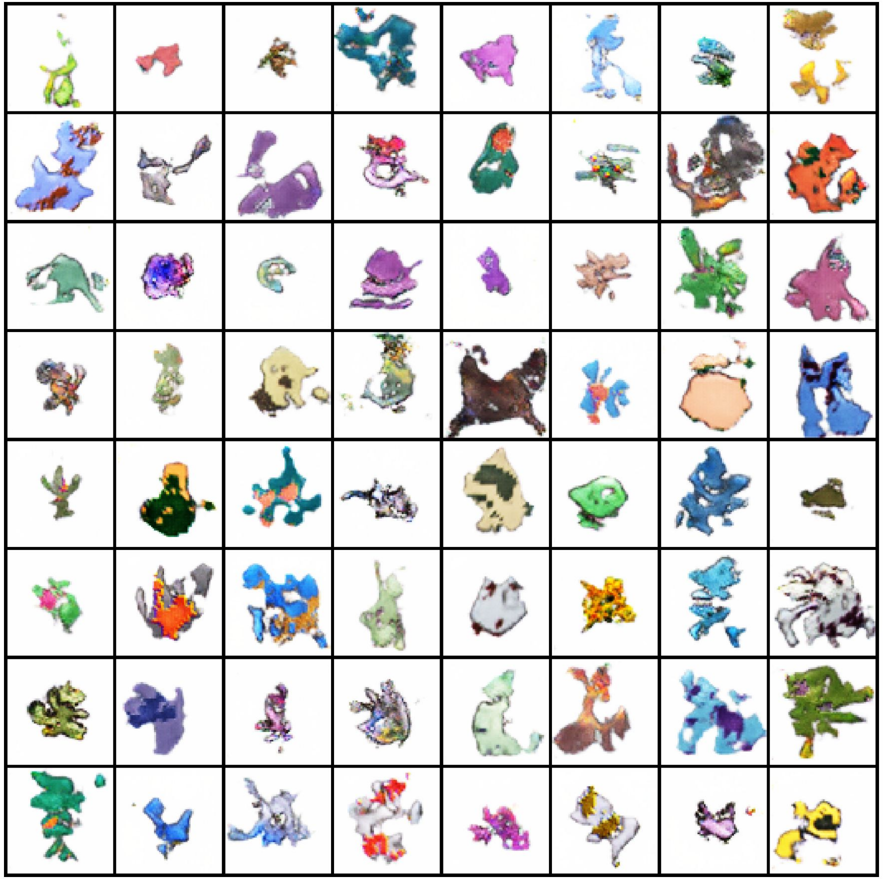
- Significant reduction in FID (-138.67!)
    - Drop from ~219 to ~80
- Better IS and Diversity
    - Higher quality, more variety
- Discriminator in a healthy range of accuracy
    - 63% : correlates with better adversarial dynamics and healthier training
    - D is good but not too good, providing useful gradients for G

Training and Validation Loss
Train Loss
Val Loss
1.4
1.2
1.0
0.8
0.6
Loss
0
20
40
60
80
100
120
Epoch


Training Metrics
Train D_Loss
Val D_Loss
Val FID
Metric Value
400
300
200
100
0
0
20
40
60
80
100
120
Epoch

Baseline Model

Final Model



Qualitative Results

- Compared to the baseline model, final model shows:
  - More coherent shapes and recognizable silhouette
  - Better color consistency (type-like palettes)
  - Cleaner edges and shading in many sprites
- Remaining issues:
  - Still somewhat blobby
  - inconsistent anatomy and symmetry
    - Eyes, limbs

Lessons Learned

- Stabilization techniques matter a lot when training GANs
  - Spectral norm, TTUR, label smoothing, early stopping
  - Good discriminator behavior correlates with better generator quality
- Dataset challenges
  - Small, highly diverse dataset (1,000 species, many poses– around 25 on average)
  - Hard for a single GAN to capture all styles and angles
- Limitations
  - Sprites lack crisp pixel-art lines
  - Inconsistent symmetry and anatomy
  - Back-view and odd angle sprites may be confusing to the model

Future Directions

- Diffusion models for sharper, more stable samples
- Class or style conditioning
  - Condition on type, color cluster, or learned cluster IDs
- Pixel-art specific approaches
  - VQ-GAN, autoregressive pixel models, or discrete convolutions
- Dataset refinements
  - Remove or separate back-view/weird-angle sprites
  - Build more curated, front-facing sprite subsets
- Longer-term
  - Interactive tool to generate and curate new “Pokémon-like” species

Thank You For Listening!