Simulating Galaxy Images via Progressive GAN

Mohamad Dia
i4DS - FHNW

Swiss Euclid Days
5 February 2020

University of Applied Sciences and Arts Northwestern Switzerland
School of Engineering

Why Simulated Images?

- Training and calibration of automated algorithms used for measurement/ detection on ver large-scale data.
    - Calibration and bias detection for shape measurement algorithms (in weak lensing) require simulated images with known ground truth lensing.
    - Training neural network classifiers (e.g. CNN) to detect strong lenses requires simulated images in order to mitigate class imbalance in the current datasets and avoid false-positive type of error.
- Simulate/generate synthetic galaxy images that mimic the real observations and exhibit real morphologies.

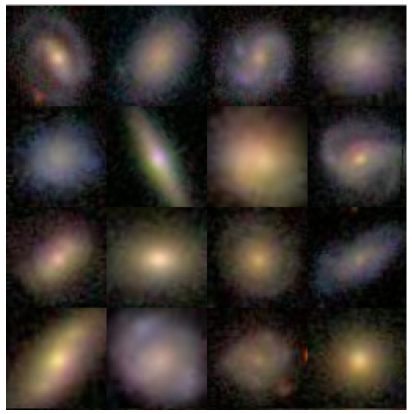
Model-Driven v.s. Data-Driven

- Model-driven (rule-based) simulation:
  - Fitting of parametric analytic profiles (size, ellipticity, brightness,…). Do not produce complex morphologies.
  - Start with high-quality galaxy images as the input of the simulation pipeline followed by a model that reproduces all the data acquisition effects. Expensive and often infeasible.

Model-Driven v.s. Data-Driven

- Model-driven (rule-based) simulation:
  - Fitting of parametric analytic profiles (size, ellipticity, brightness,…). Do not produce complex morphologies.
  - Start with high-quality galaxy images as the input of the simulation pipeline followed by a model that reproduces all the data acquisition effects. Expensive and often infeasible.
- Data-driven simulation (generative models in ML):
  - Generate new data with similar density distribution as the existing training data (unsupervised learning).

Generative Models



Training data*

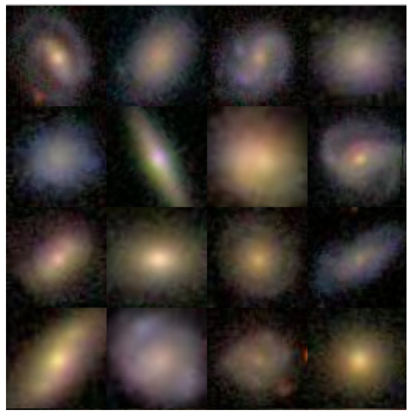
New synthetic sample

$$x \sim P_{\mathrm{data}}(x)$$

$$x \sim P_{\mathrm{model}}(x) \approx P_{\mathrm{data}}(x)$$

* Galaxy-Zoo dataset (SDSS).

Generative Models



Training data*

New synthetic sample

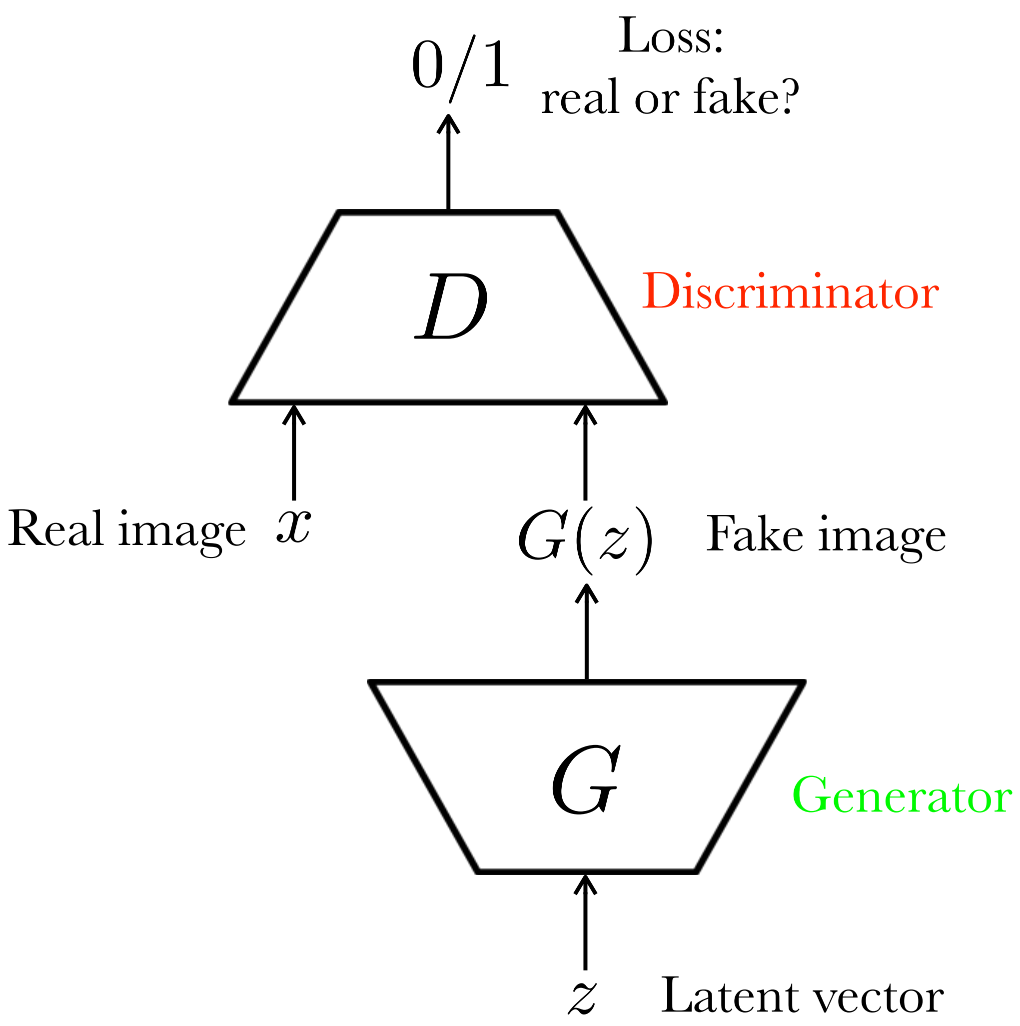
$x \sim P_{\text{data}}(x)$ $\qquad$ $x \sim P_{\text{model}}(x) \approx P_{\text{data}}(x)$

- Explicit density estimation:
  - Classical probabilistic models (e.g. GMM).
  - Neural network approach: Pixel-CNN, Pixel-RNN, VAE, …
- Implicit density estimation:
  - Generative Adversarial Network GAN.

* Galaxy-Zoo dataset (SDSS).

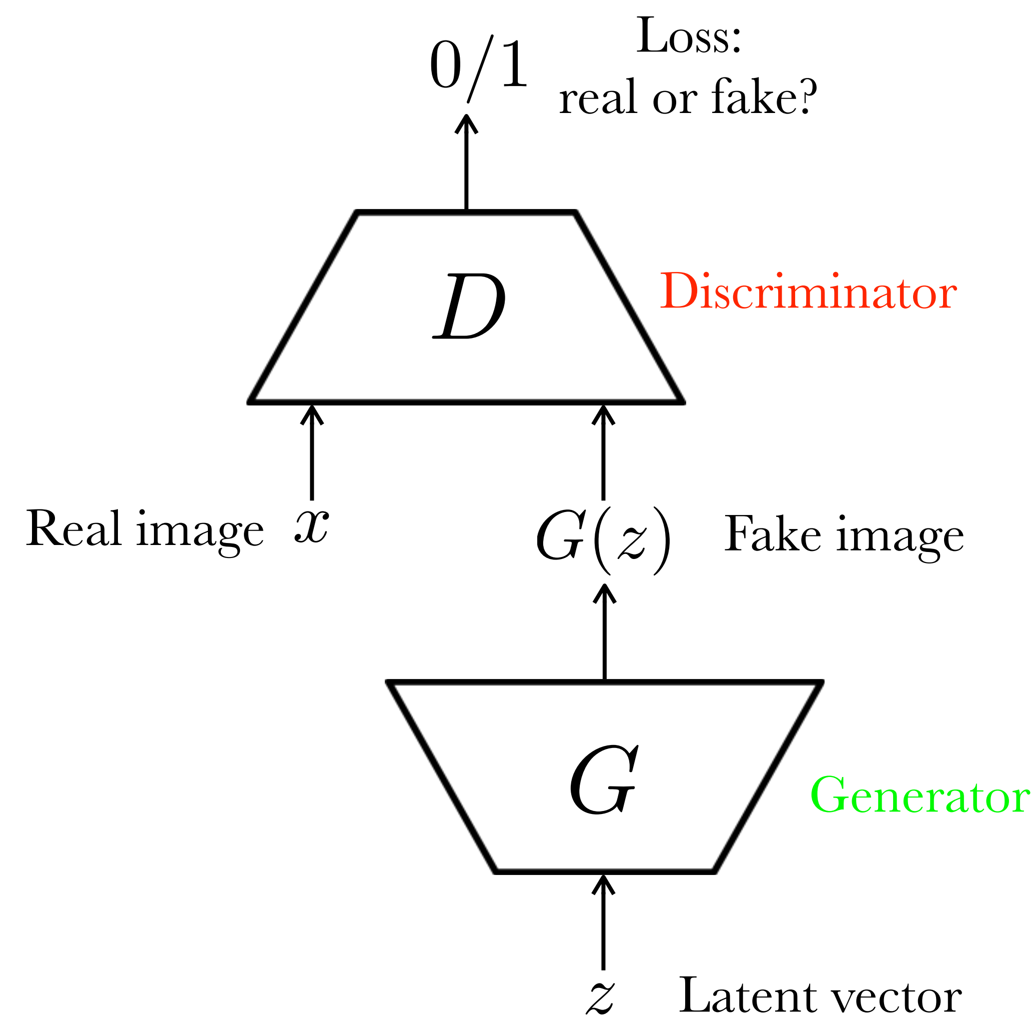
Generative Adversarial Network

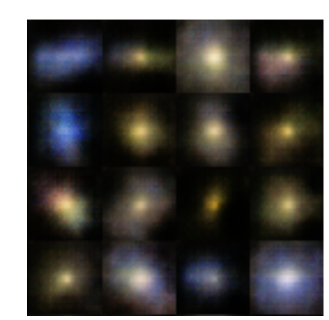
- Two-player minimax game:
  - Two competing neural nets (mirrored CNN architectures).
  - Generator samples simulated (fake) images from random latent space.
  - Discriminator plays the role of adaptive loss.
  - Joint optimization (alternate training between generator and discriminator).
  - Standard version hard to train (saddle point equilibrium): unstable behaviour, convergence issues, mode collapse, etc…



Generative Adversarial Network

- Two-player minimax game:
  - Two competing neural nets (mirrored CNN architectures).
  - Generator samples simulated (fake) images from random latent space.
  - Discriminator plays the role of adaptive loss.
  - Joint optimization (alternate training between generator and discriminator).
  - Standard version hard to train (saddle point equilibrium): unstable behaviour, convergence issues, mode collapse, etc…

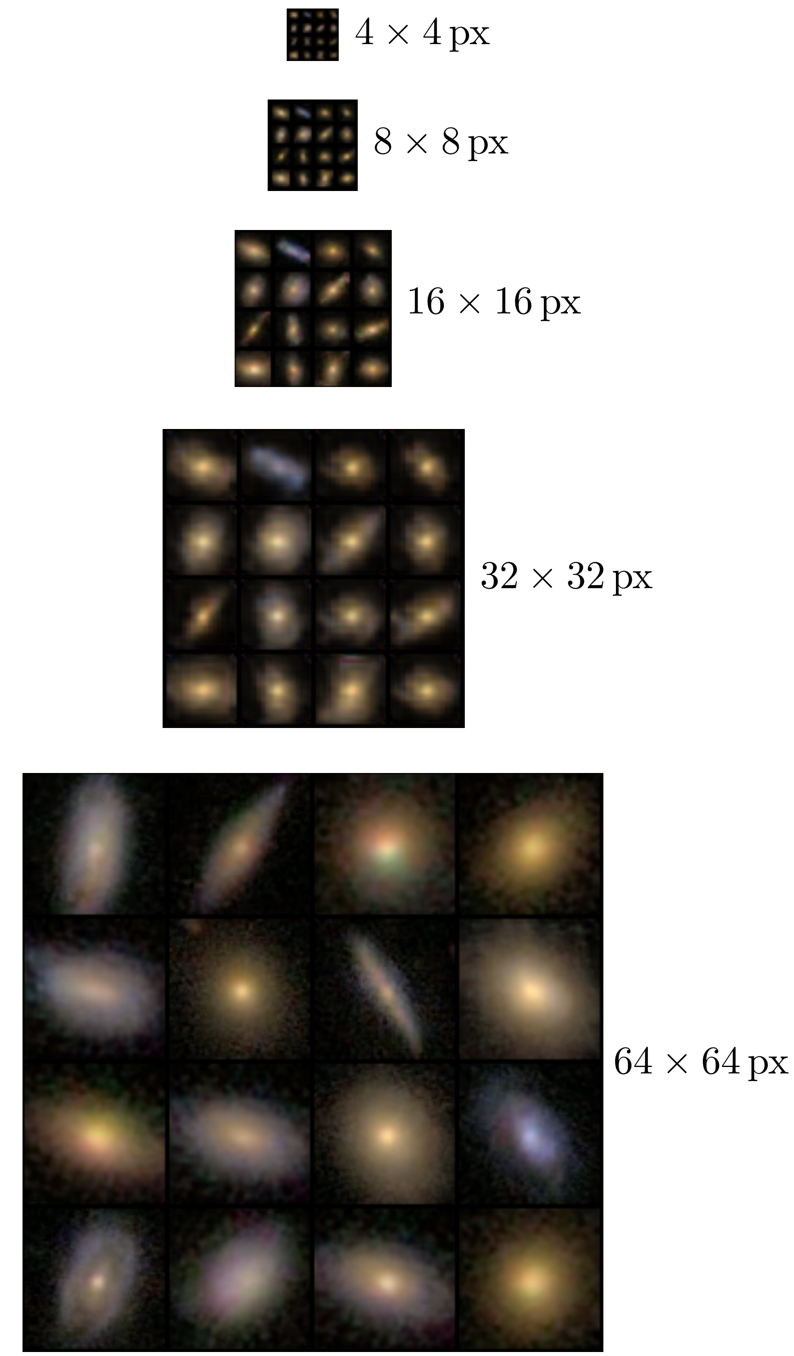




Improved Training of GAN

- Progressive training methodology:
  - Start with low resolution (easy to train) and keep on growing both networks by **adding layers synchronously and smoothly** [Karras et al. 2018].

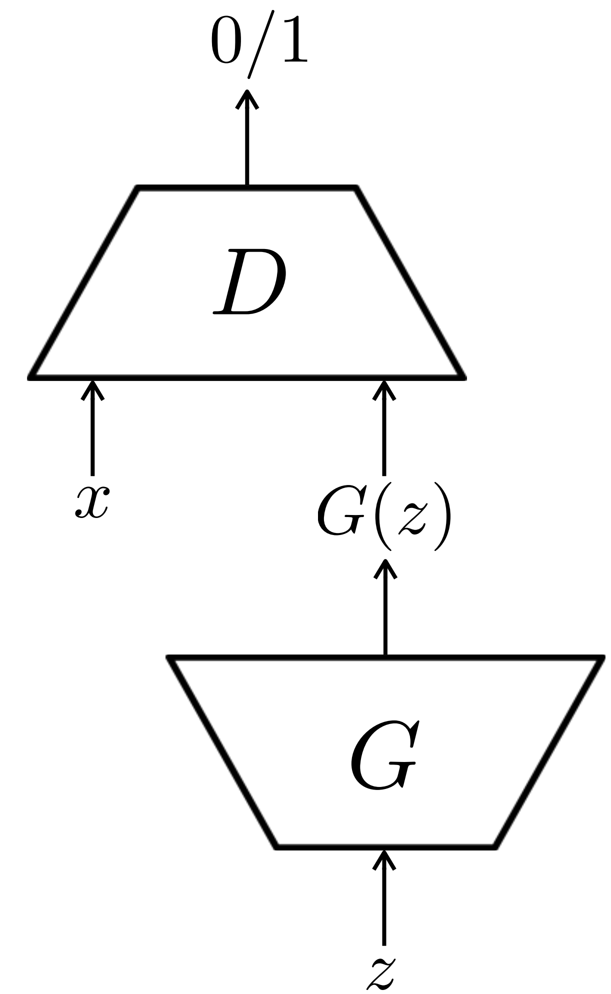
- Stable loss function :
  - **Wasserstein distance** with gradient penalty to mitigate vanishing gradient [Arjovsky et al. 2017].



[Dia et al. 2019]

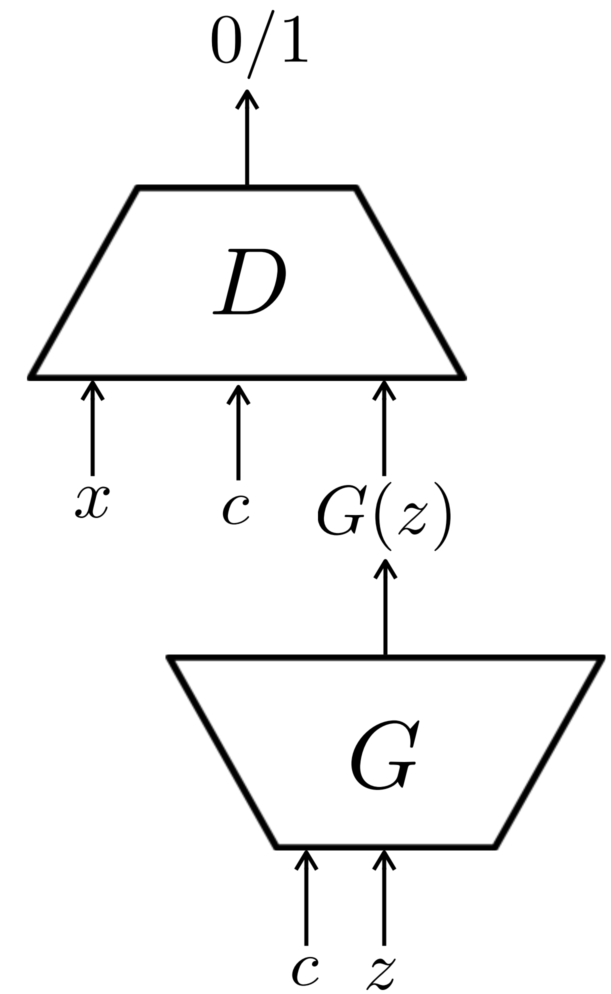
Possible Extensions

- Using labels (when available):
    - Control the generation task.
    - Improve the quality.
    - Increase the diversity.
    - Conditional GAN: learn a model that samples from a conditional distribution.
    - Info GAN, Auxiliary Classifier GAN.



Possible Extensions

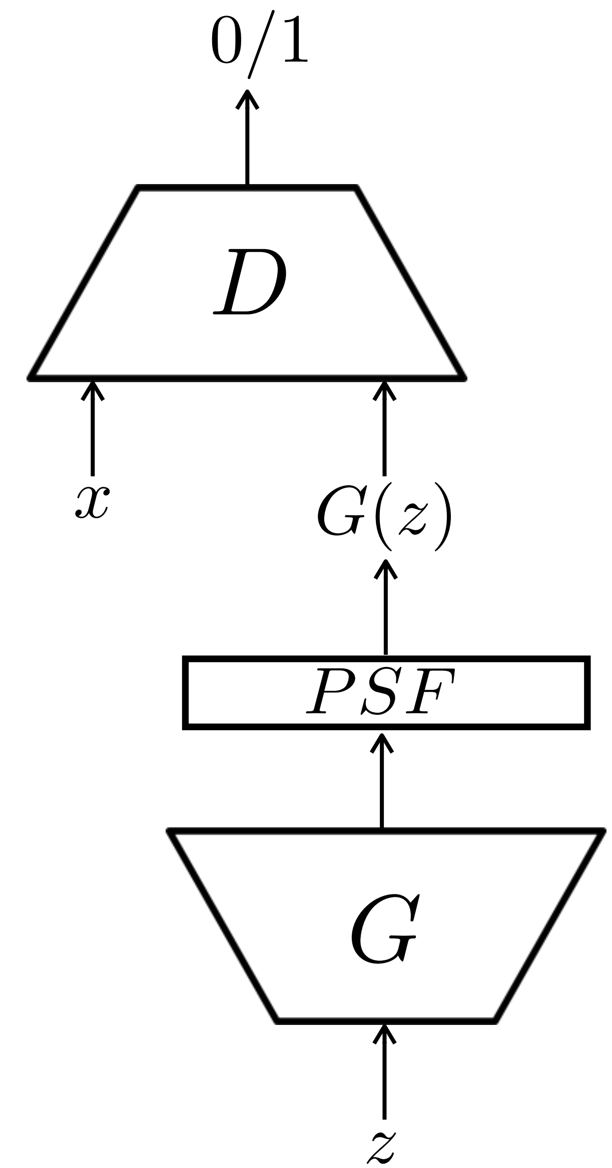
- Using labels (when available):
  - Control the generation task.
  - Improve the quality.
  - Increase the diversity.
  - Conditional GAN: learn a model that samples from a conditional distribution.
  - Info GAN, Auxiliary Classifier GAN.



$$x \sim P_{\text{model}}(x|c) \approx P_{\text{data}}(x|c)$$

Possible Extensions

- Generate PSF deconvolved images:
  - The training dataset is PSF convolved (for certain instrument, i.e. DES, SDSS,…).
  - Add a model-based PSF convolution layer to the generator during training.
  - The PSF layer is non-trainable.
  - Get access to PSF deconvolved images during generation (independent of the instrument).
  - Convolve the images with another instrument's PSF model.



THANK YOU