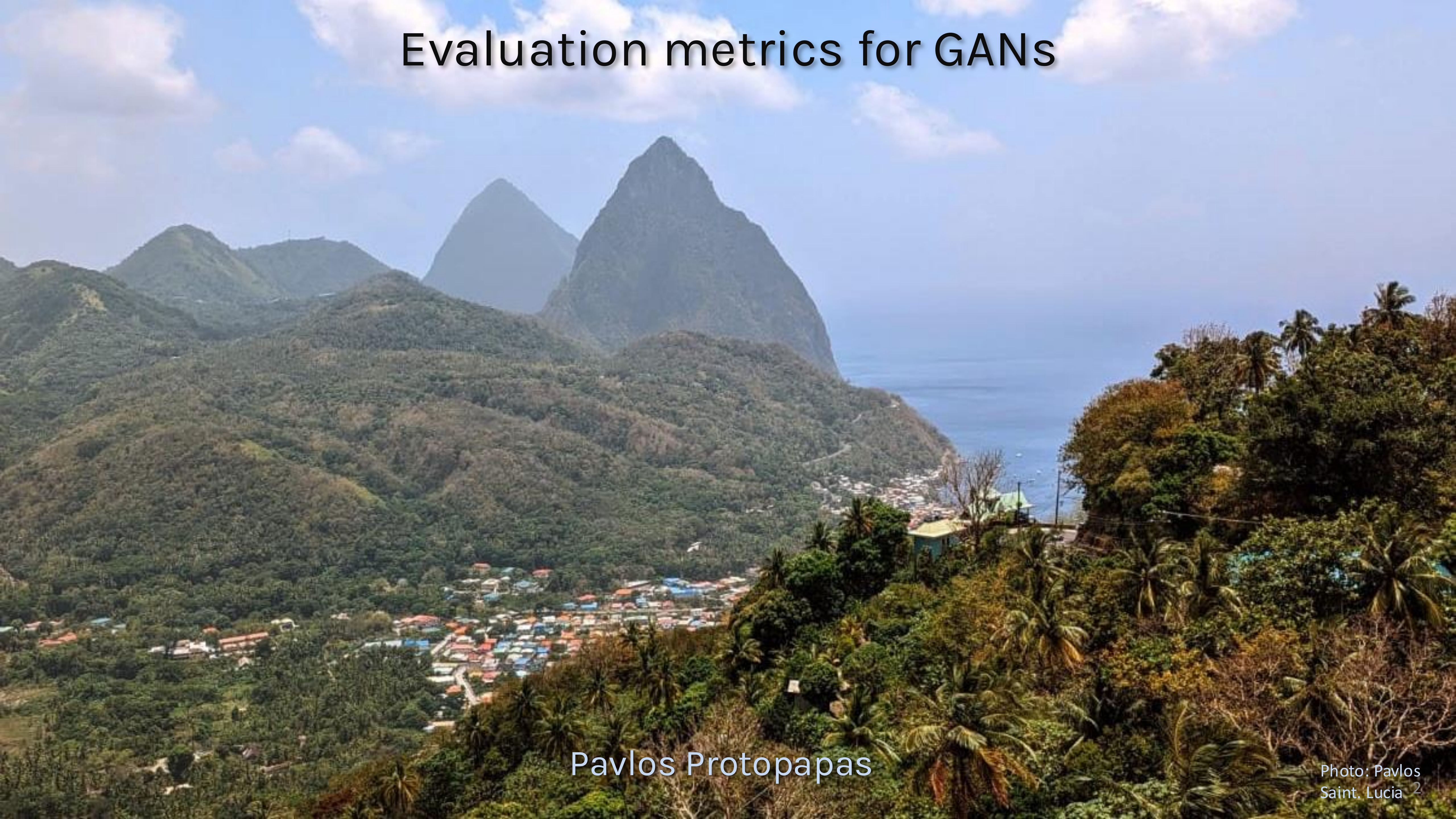
Evaluation metrics for GANs



Pavlos Protopapas

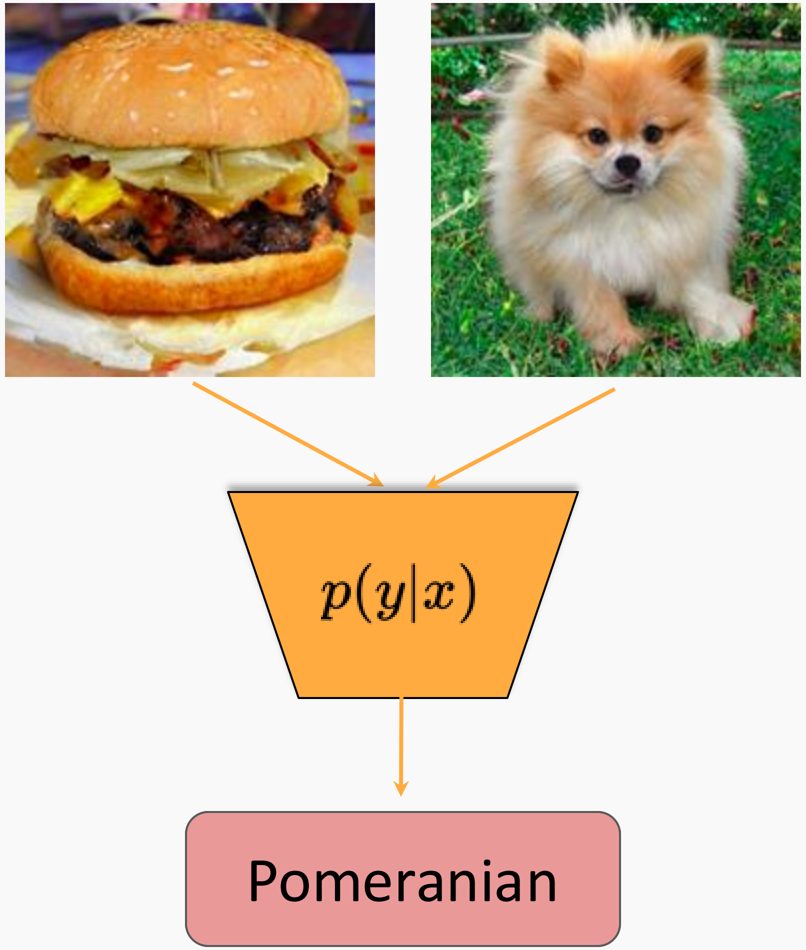
Photo: Pavlos
Saint. Lucia

Outline

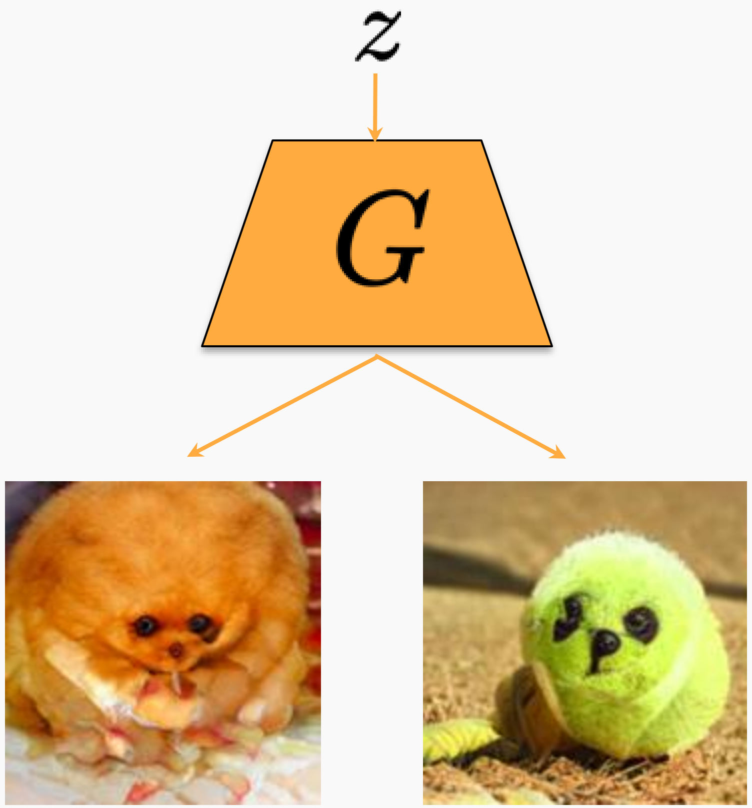
- **Evaluation Metrics for GANs**
  - Inception score
  - Train Synthetic Test Real
  - Fréchet Distance

Evaluating GANs: Why is it challenging?

Supervised Learning



GANs



??

~~Could we use D?~~

Evaluating GANs: What do we want from a Generator?

Fidelity: Quality of images

Diversity: Variety of images

[A Google intern built the AI behind these shockingly good fake images]

Evaluating GANs

- Human Annotators
- Inception score
- Train Synthetic Test Real
- Fréchet Distance

And more..

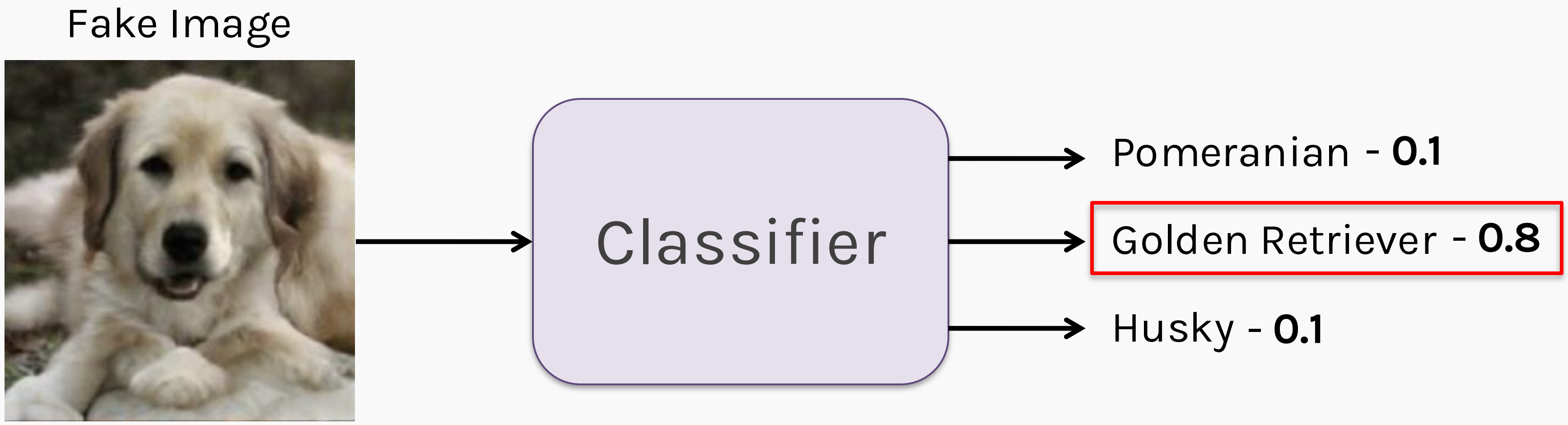
Outline

- Evaluation Metrics for GANs
  - **Inception score**
  - Train Synthetic Test Real
  - Fréchet Distance
- Challenges in GANs

Evaluating GANs: Inception Score

Consider a generative model that generates images of dogs and another pre-trained classifier that is trained on real images to classify different types of dog breeds.

In the ideal case:



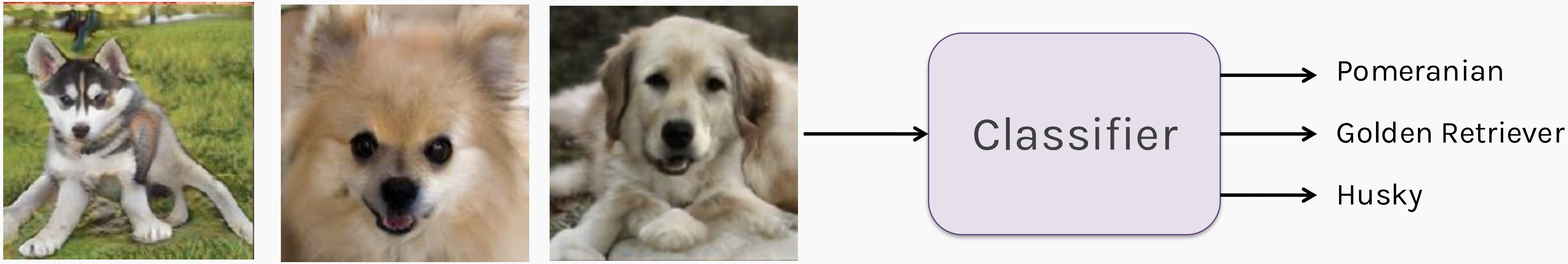
$p(y|x)$ should be very narrow (high for one class, low for other classes) for all images since there should be no uncertainty when classifying.
This indicates high fidelity of images.

Evaluating GANs: Inception Score

Consider a generative model that generates images of dogs and another pre-trained classifier that is trained on real images to classify different types of dog breeds.

In the ideal case:

Fake Images



We could expect a uniform distribution for all classes → $p(y)$ should be very wide.
This indicates high diversity of images.

Evaluating GANs: Inception Score

Therefore,

For high fidelity: $p(y|x)$ should be very narrow.

For high diversity: $p(y)$ should be very wide.

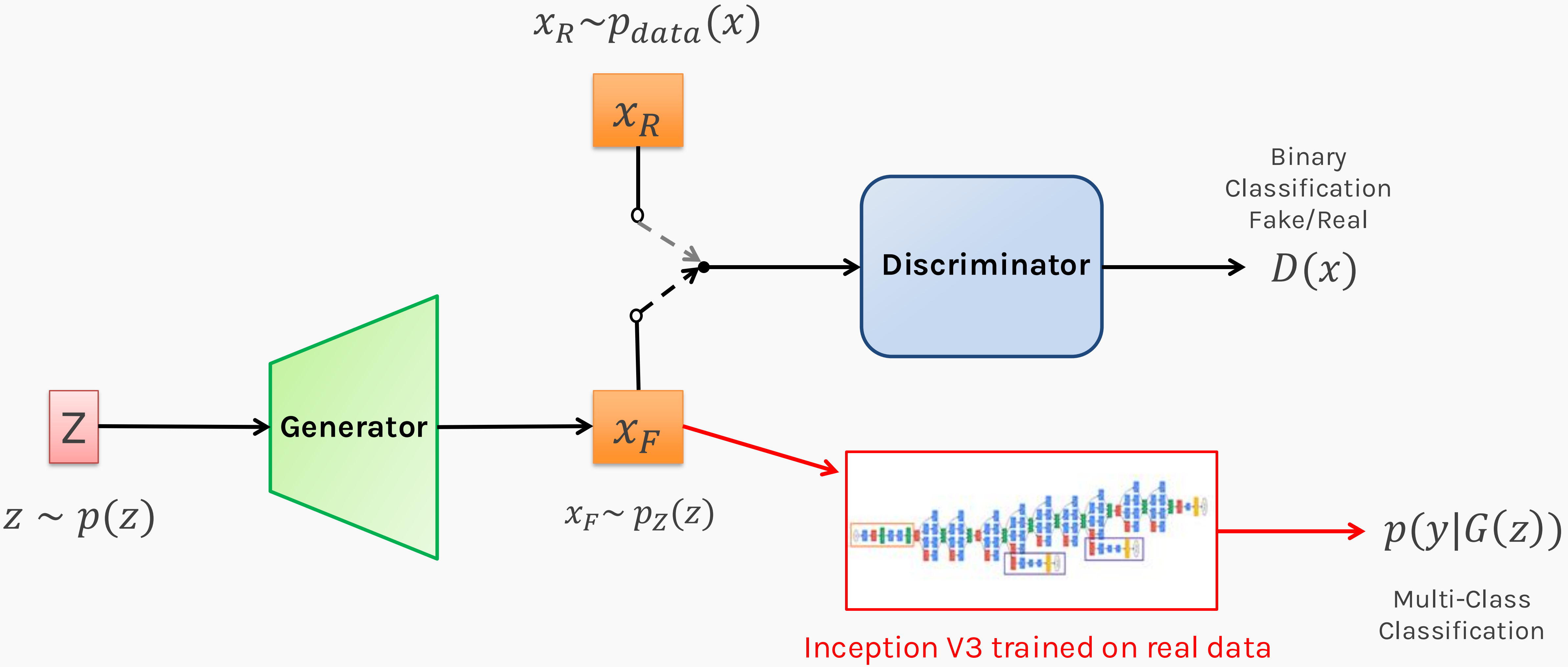
Therefore, $\boldsymbol{p(y)}$ and $\boldsymbol{p(y|x)}$ should be very different!

Evaluating GANs: Inception Score

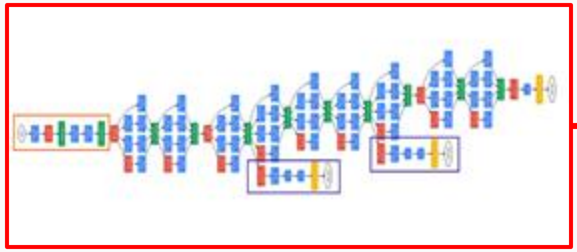
In order to do this, we need a good classifier.

In the context of images, we can use the inception network and call it inception score.

Evaluating GANs: Inception Score



Evaluating GANs: Inception Score

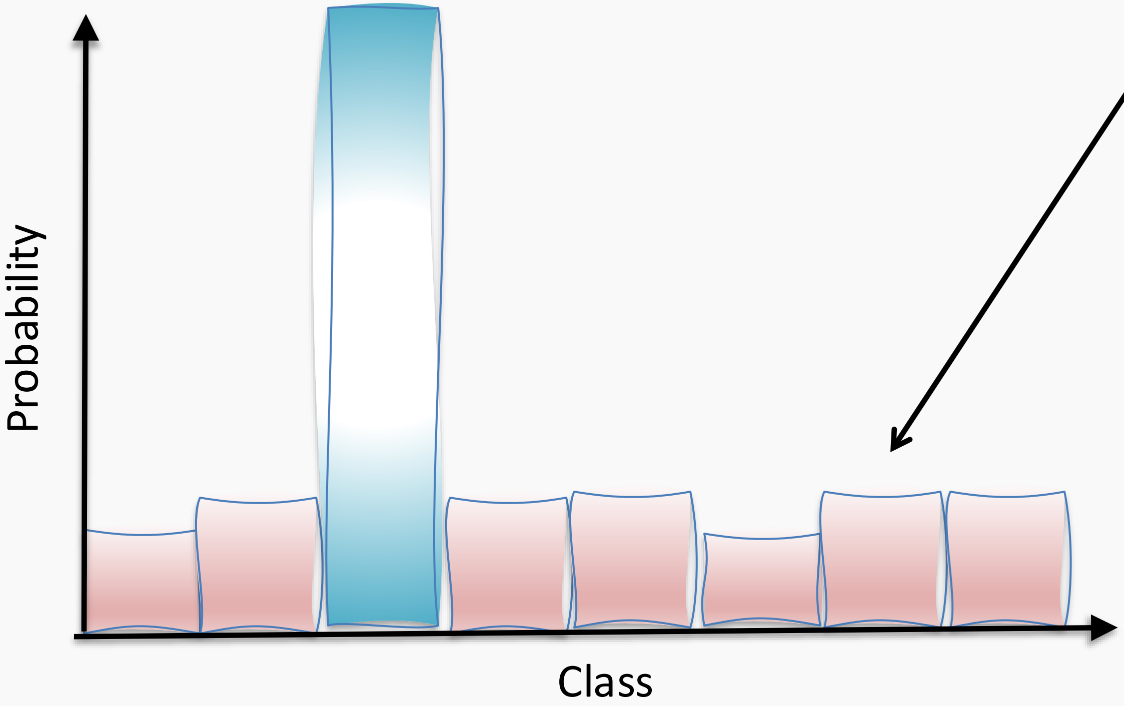


$$p(y|G(z))$$

Should be narrow indicating high quality (low entropy).

$$p(y) = \int p(y|G(z))dz$$

Distribution of labels which should be uniform indicating high diversity (high entropy).



Evaluating GANs: Inception Score

$$Inception\ Score\ (G) = \exp(\mathbb{E}_{x_F \sim p_g} D_{KL}(\,p(y|x_F)\ ||\ p(y)))$$

Exponent is used to scale the KL divergence to a readable score.
Eg. 100 for Inception Score and not 0.000001.

KL measures the difference of the two distributions.

Therefore, if we want high quality and diverse images, KL has to be high.

Evaluating GANs: Inception Score

Inception score (IS) is used for evaluation not for training.

Using IS for training, yields weird results.

Figure 1. Sample of generated images achieving an Inception Score of 900.15. The maximum achievable Inception Score is 1000, and the highest achieved in the literature is on the order of 10.

Evaluating GANs: Inception Score

Disadvantages of Inception Score:

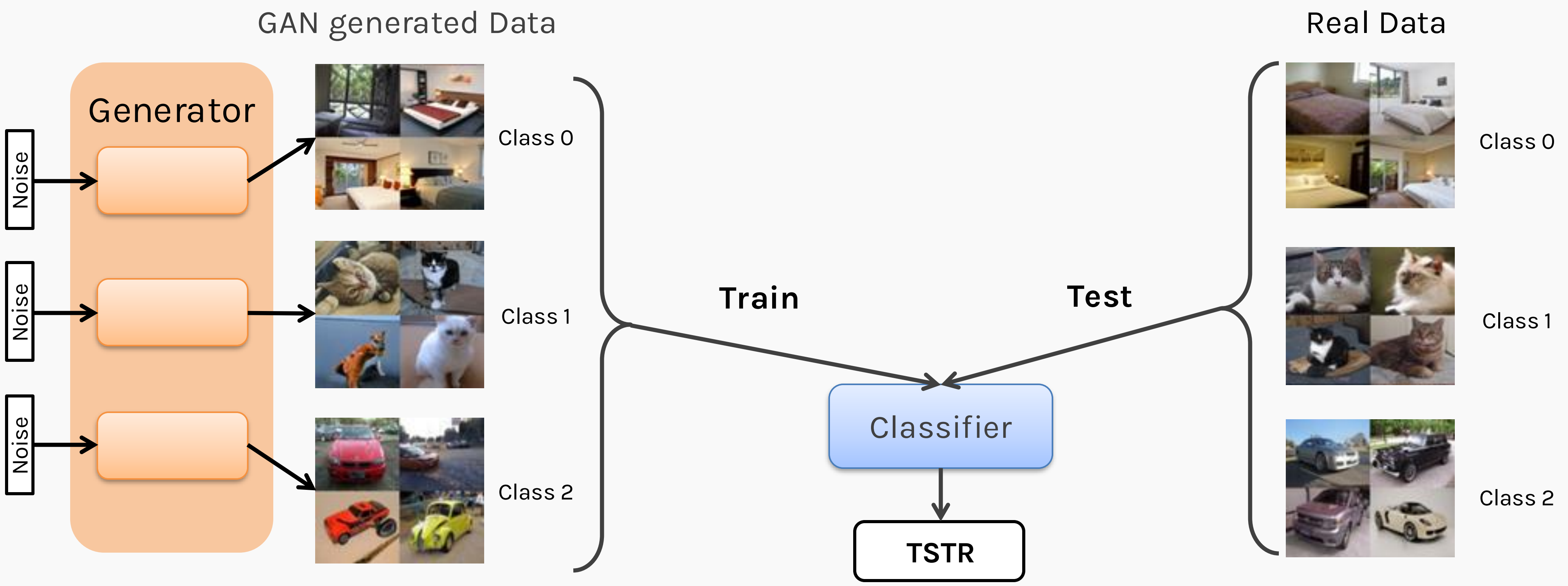
- If your generator generates only one image per class, repeating each image many times, it can score highly.
- Depends on the classifier's classes. For example, if one image contains 2 or more objects of different classes, the score is reduced because of high entropy in several classes.

Outline

- Evaluation Metrics for GANs
  - Inception score
  - **Train Synthetic Test Real**
  - Fréchet Distance
- Challenges in GANs

Evaluating GANs: Train Synthetic, Test Real: TSTR

After training the GAN, we train another classifier on generated data and test the classifier on real data.



If synthetic data are of high quality the we expect TSTR≥ TRTR (Train Real, Test Real).

Outline

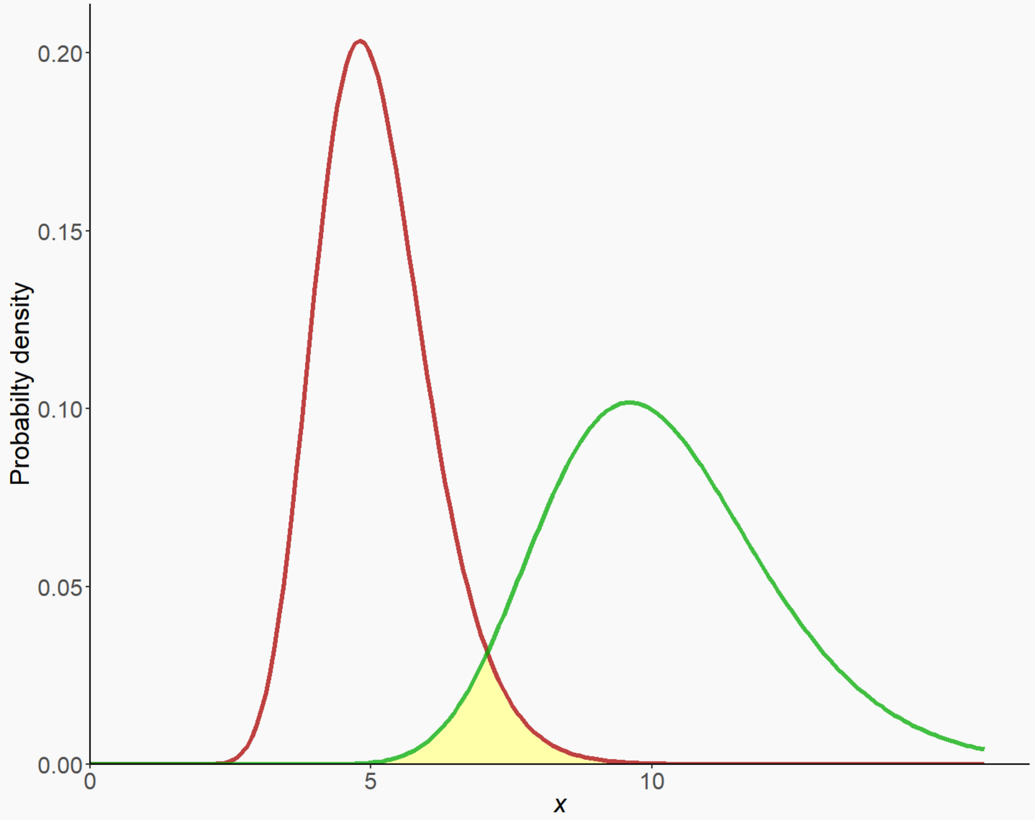
- Evaluation Metrics for GANs
  - Inception score
  - Train Synthetic Test Real
  - **Fréchet Distance**
- Challenges in GANs

Evaluating GANs: Fréchet Inception Distance

The Fréchet distance is a measure of similarity between two distributions.

The “distance” between two 1-D normal distributions is:

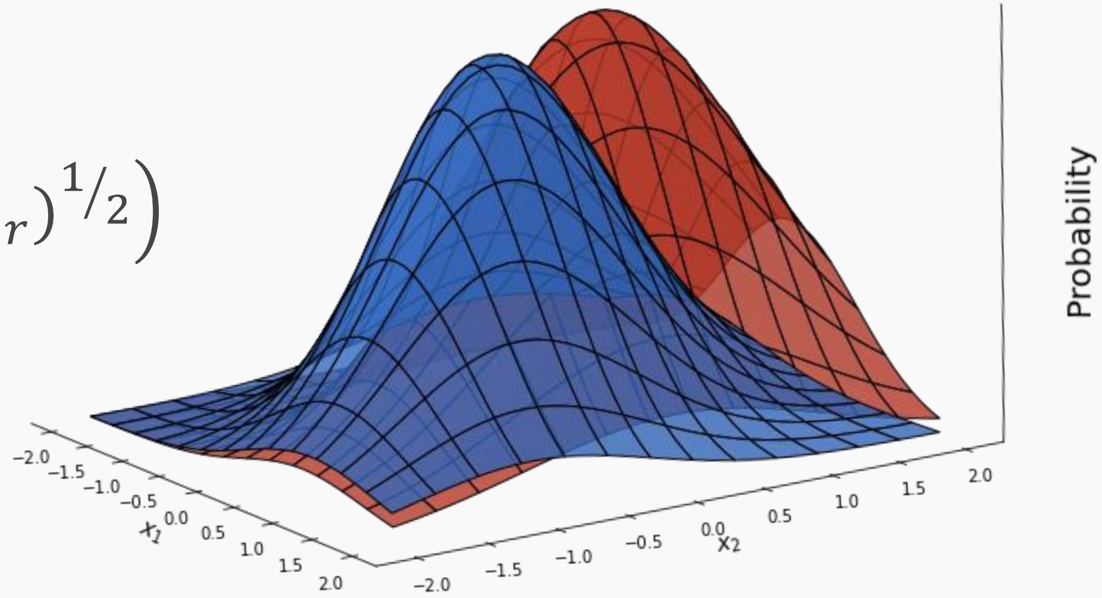
$$d(X_r, X_g) = (\mu_r - \mu_g)^2 + (\sigma_r - \sigma_g)^2$$



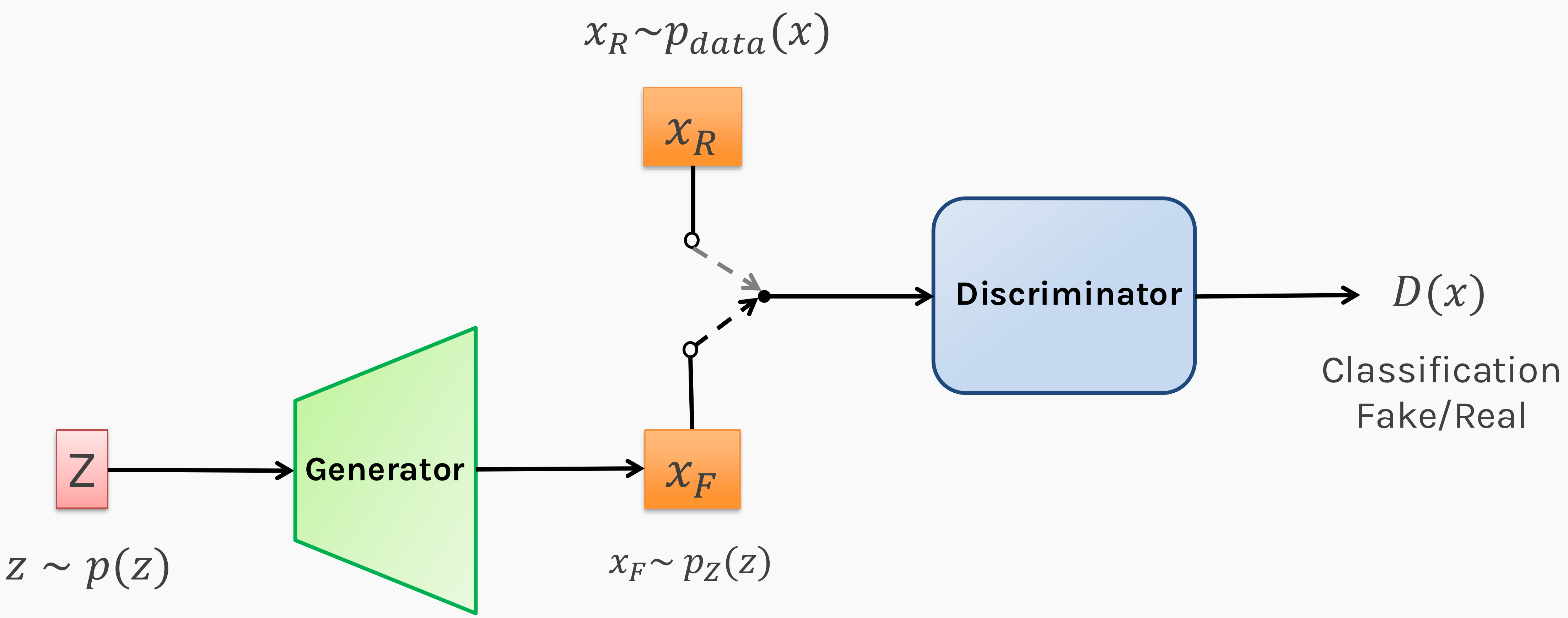
Evaluating GANs: Fréchet Inception Distance

The “distance” between two multi-variate normal distributions is:

$$d(X_r, X_g) = |\mu_r - \mu_g|^2 + Tr\left(\Sigma_r + \Sigma_g - 2(\Sigma_g\Sigma_r)^{1/2}\right)$$



Evaluating GANs: Fréchet Inception Distance

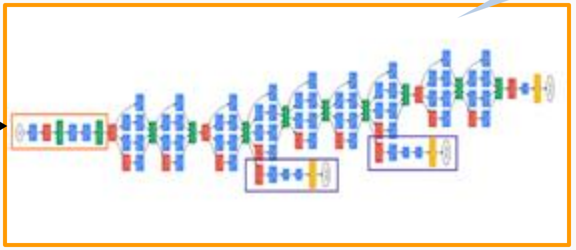


Evaluating GANs: Fréchet Inception Distance

Not always InceptionV3, it can be any classifier.

$x_R \sim p_{data}(x)$

x_R

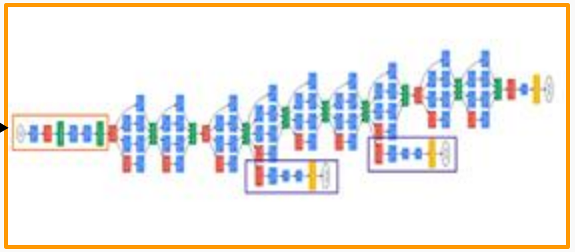


Inception V3

Z

Generator

x_F



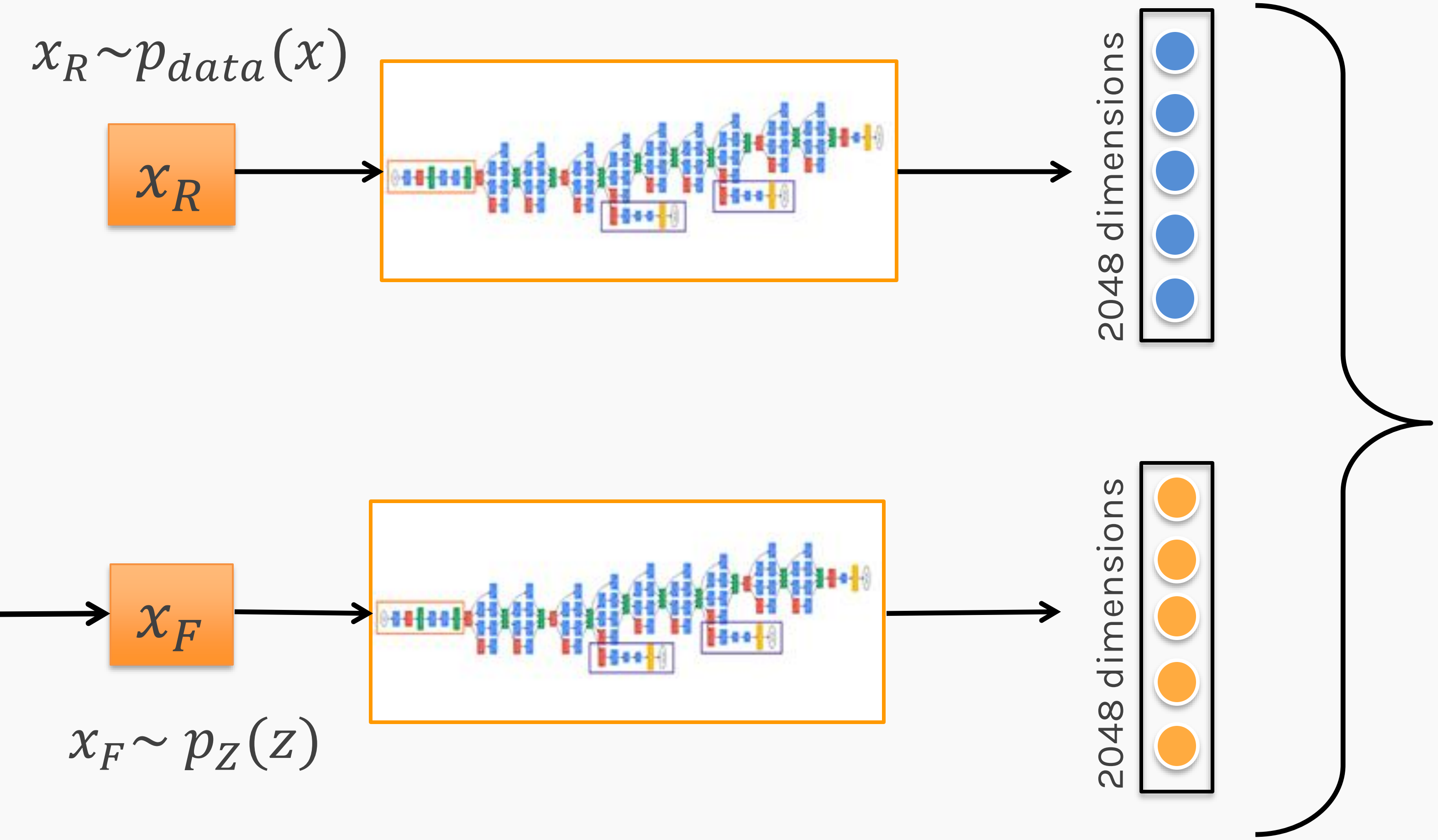
Inception V3

$z \sim p(z)$

$x_F \sim p_Z(z)$

type	patch size/stride or remarks	input size
conv	3×3/2	299×299×3
conv	3×3/1	149×149×32
conv padded	3×3/1	147×147×32
pool	3×3/2	147×147×64
conv	3×3/1	73×73×64
conv	3×3/2	71×71×80
conv	3×3/1	35×35×192
3×Inception	As in figure 5	35×35×288
5×Inception	As in figure 6	17×17×768
2×Inception	As in figure 7	8×8×1280
pool	8 × 8	8 × 8 × 2048
linear	logits	1 × 1 × 2048
softmax	classifier	1 × 1 × 1000

Evaluating GANs: Fréchet Inception Distance

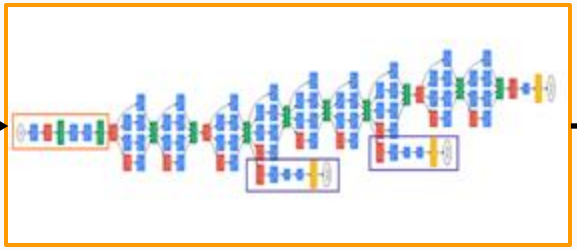


Assuming that the embeddings distribute as multivariate gaussians, compute the FID distance between the distributions.

Evaluating GANs: Fréchet Inception Distance

$x_R \sim p_{data}(x)$

x_R



2048 dimensions

$\sim N(\mu_r, \Sigma_r)$

x_F

$x_F = G(z)$

2048 dimensions

$\sim N(\mu_g, \Sigma_g)$

$$FID = |\mu_r - \mu_g|^2 + Tr\left(\Sigma_r + \Sigma_g - 2(\Sigma_g \Sigma_r)^{1/2}\right)$$

In practice, we use large batches, take μ_g, Σ_g and the average the μ_g, Σ_g's over batches.

Evaluating GANs: Fréchet Inception Distance

Disadvantages of FID:

- There is no interpretable range for FID.
- The FID score changes depending on the number of images you choose to sample from the generator. As the number of samples increase, the FID score decreases.
- It can be slow to run depending on the dimensionality of the embedding and the sample size.
- We assume that the distribution is multivariate normal, and the distribution is completely defined by mean and covariance.