The Fault in Our Metrics: Revisiting Generative Model Evaluation with Chamfer Distance

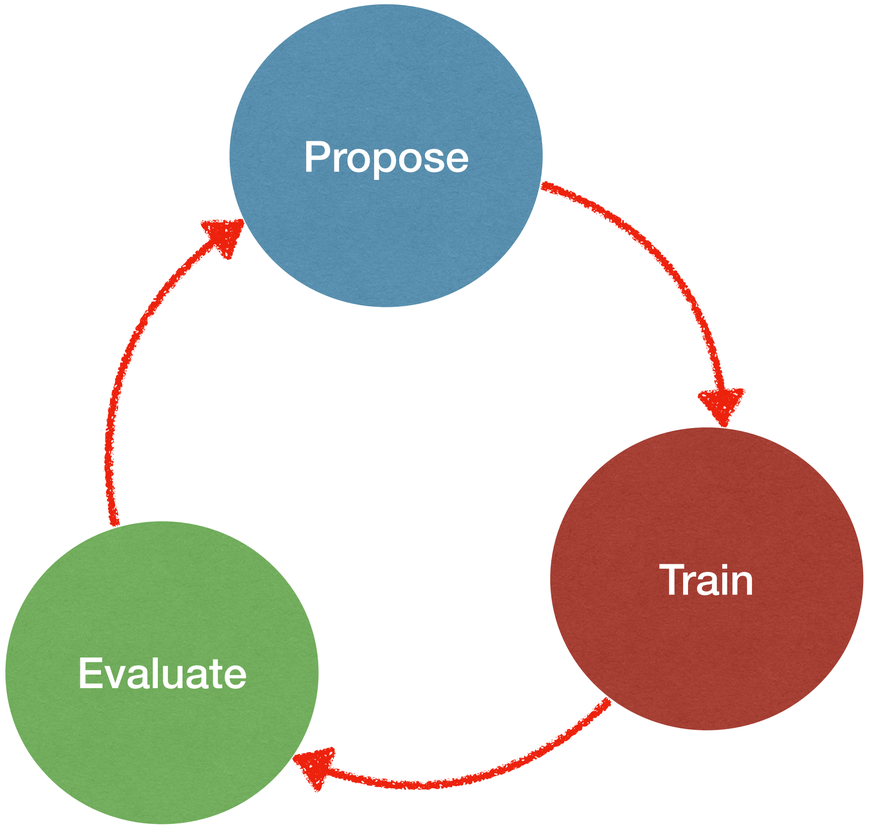
TLDR: Existing evaluation metrics for generative models are fundamentally flawed; we introduce a simple, principled replacement based on nearest-neighbour distances.

Chirag Vashist, *Vaibhav Santhosh, Tristan Engst, Yanshu Zhang, Ke Li*

Motivation

Image generation models are judged on thousands of samples, so we rely on automated metrics.

Metrics decide which methods win and where the field goes.



In this paper, we ask the question:

Are evaluation metrics informative or reliable enough ?

Roadmap

1. **Distributional Metrics**
    1.1. Examples: FID, KID, CMMD
    1.2. Problems with a symmetric score
2. **Disentangled Metrics**
    2.1. Formulation: ball-based metrics
    2.2. Problems: inflated balls and binary tests
3. **Solution: Chamfer Distance**
    3.1. Intuition and formulation
    3.2. Experiments: sensitivity to image distortions

Distributional Metrics

Definition

Real datapoints Generated samples

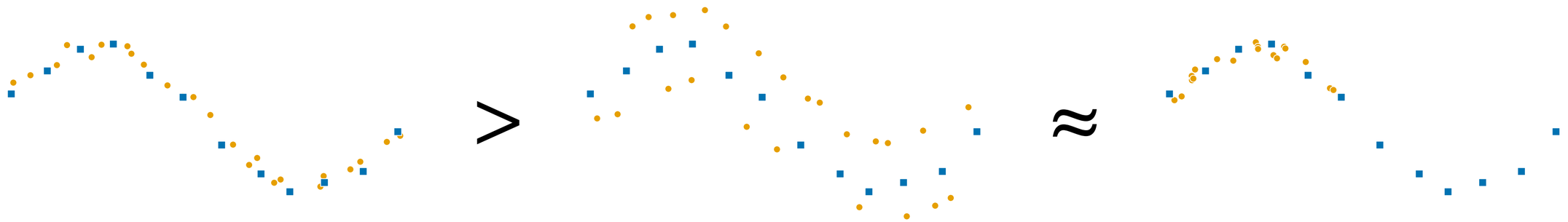
Distributional metrics return a single, symmetric score by comparing the real and generated sets in a feature space of a encoder.

Examples

1. FID: fits a Gaussian to each set of Inception-v3 features and computes the Fréchet distance between the two Gaussians.
2. FD_{DINOv2}: the same as FID, but with DINOv2 features.
3. KID, CMMD: compare the two sets with Maximum Mean Discrepancy (MMD), so they make no Gaussian assumption.

Problem

These metrics are symmetric: d(P, Q) = d(Q, P). The score tells you how large the gap is, but not where it comes from: poor fidelity or poor diversity.



High fidelity and coverage High coverage, low fidelity High fidelity, low coverage

Distributional Metrics

Problems

BigGAN: FID 6.67

IMM (1 step): FID 8.11

Jack o'-lantern





Daisy





Samples from IMM look more visually diverse, yet FID ranks BigGAN better than IMM

Disentangled Metrics

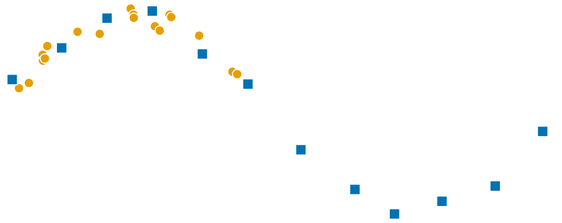
Intuition of KNN-ball based metrics

Disentangled metrics address this limitation by proposing metrics that express the performance of a model along two separate axes: **Fidelity** and **Coverage**

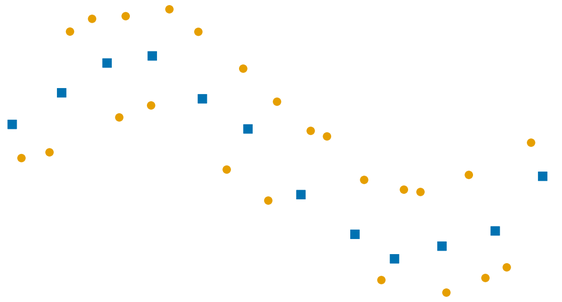
The most notable of existing ones are **Precision** and **Recall**, which approximate the local manifold of each distribution by performing nearest-neighbour search among its samples and using the resulting distances to construct isotropic balls in feature space.

Evaluation is then performed by checking whether samples from the other distribution fall within this approximated manifold.

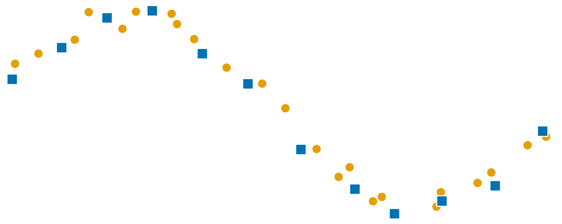
Followups include: **Density/Coverage** and **Symmetric Precision/Recall**



High fidelity, low coverage



High coverage, low fidelity



High fidelity and coverage

Real datapoints Generated samples

KNN-ball based metrics

Definition: Precision

Let P and Q be the real and candidate distributions respectively :

$\{x_i\}_{i=1}^{N} \sim P$ and $\{y_j\}_{j=1}^{M} \sim Q$

1. Estimate support of P using union of balls centred at each real sample:

$$\widehat{\mathscr{M}}_X := \bigcup_{i=1}^{N} B\Big(f(x_i), \mathrm{NND}_k\left(f(x_i)\right)\Big)$$

2. Define Precision through binary membership:

$$\text{Precision} := \frac{1}{M}\sum_{j=1}^{M} \mathbf{1}\Big[f(y_j) \in \widehat{\mathscr{M}}_X\Big]$$



KNN-ball based metrics

Definition: Precision

Let P and Q be the real and candidate distributions respectively :

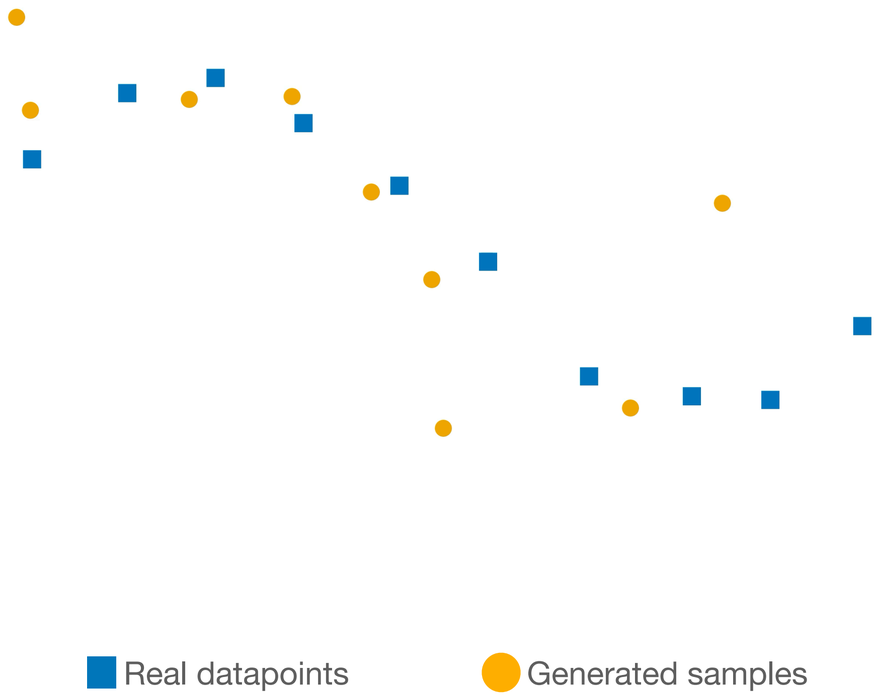
$\{x_i\}_{i=1}^{N} \sim P$ and $\{y_j\}_{j=1}^{M} \sim Q$

1. Estimate support of P using union of balls centred at each real sample:

$$\widehat{\mathscr{M}}_X := \bigcup_{i=1}^{N} B\Big(f(x_i), \mathrm{NND}_k\left(f(x_i)\right)\Big)$$

2. Define Precision through binary membership:

$$\text{Precision} := \frac{1}{M}\sum_{j=1}^{M} \mathbf{1}\Big[f(y_j) \in \widehat{\mathscr{M}}_X\Big]$$



KNN-ball based metrics

Definition: Precision

Let P and Q be the real and candidate distributions respectively :

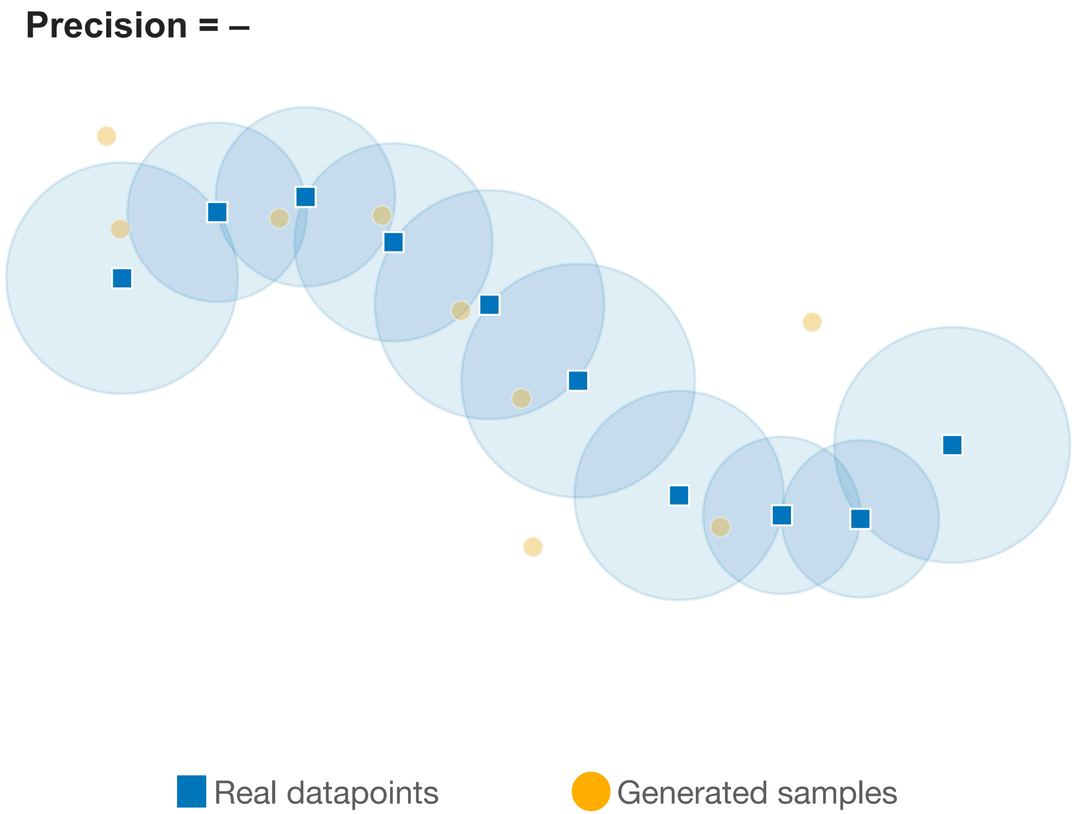
$\{x_i\}_{i=1}^{N} \sim P$ and $\{y_j\}_{j=1}^{M} \sim Q$

1. Estimate support of P using union of balls centred at each real sample:

$$\widehat{\mathscr{M}}_X := \bigcup_{i=1}^{N} B\Big(f(x_i), \mathrm{NND}_k\left(f(x_i)\right)\Big)$$

2. Define Precision through binary membership:

$$\text{Precision} := \frac{1}{M}\sum_{j=1}^{M} \mathbf{1}\Big[f(y_j) \in \widehat{\mathscr{M}}_X\Big]$$



KNN-ball based metrics

Definition: Recall

Let P and Q be the real and candidate distributions respectively :

$\{x_i\}_{i=1}^{N} \sim P$ and $\{y_j\}_{j=1}^{M} \sim Q$

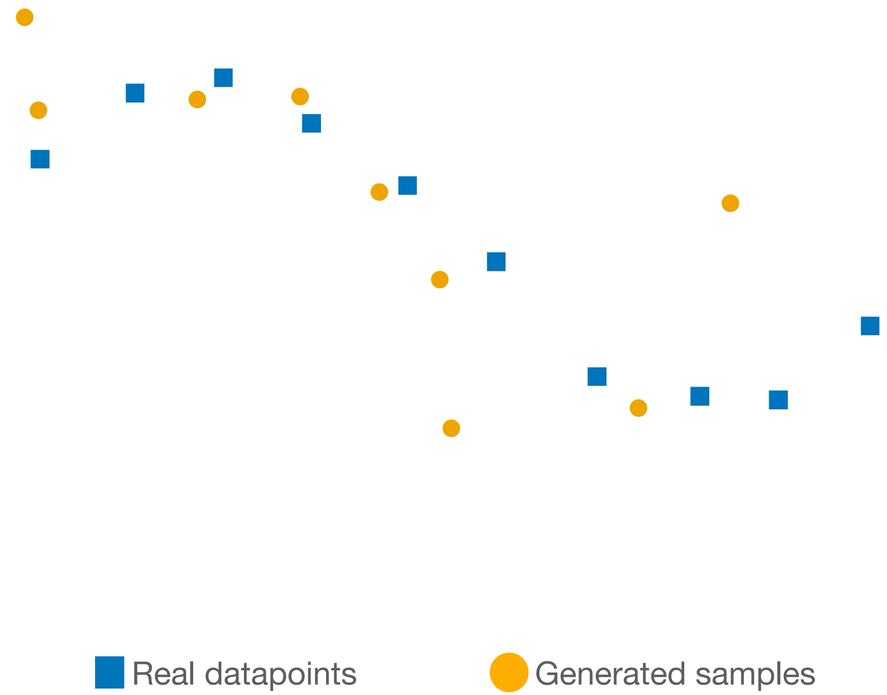
1. Estimate support of Q using union of balls centred at each generated sample:

$$\widehat{\mathscr{M}}_Y := \bigcup_{j=1}^{M} B\left(f(y_j), \mathrm{NND}_k\left(f(y_j)\right)\right)$$

2. Define Recall through binary membership:

$$\text{Recall} := \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[f(y_i) \in \widehat{\mathscr{M}}_Y\right]$$

Recall = –



KNN-ball based metrics

Problems: Volume of balls can be high; binary test can be brittle

Metric	Value
Precision ↑	1.0
Recall ↑	1.0
Density ↑	2.5
Coverage ↑	1.0
Sym. Precision ↑	1.0
Sym. Recall ↑	1.0
FD ↓	0.0

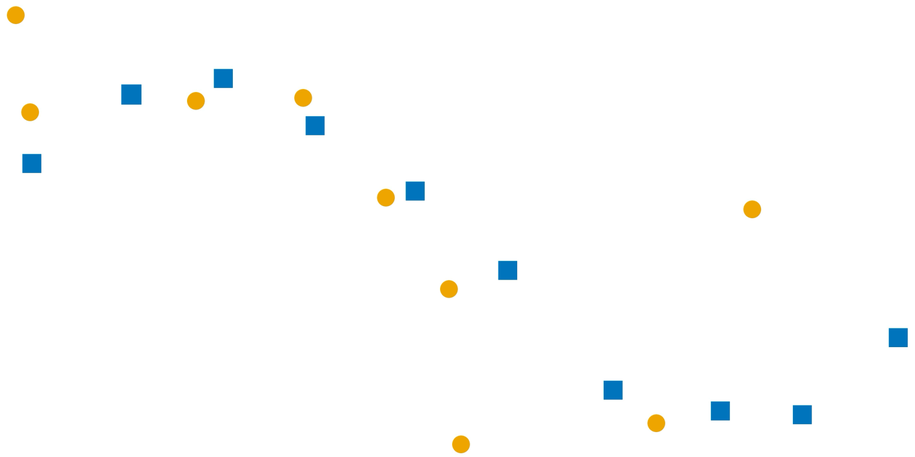
The two distributions are clearly distinct; yet they get perfect scores !



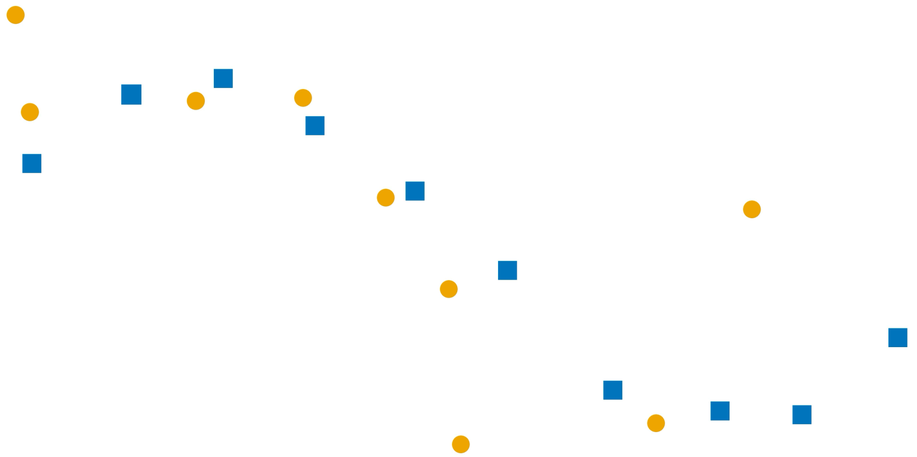
Proposed Metric

Intuition: Drop the ball; keep the distance

KNN-ball based metrics: NN search *within* set



Proposed Intuition: NN search *across* sets



Real datapoints Generated samples

Proposed Metric: Fidelity

Fake

Real Nearest Neighbours

Increasing distance





d = 4.70



d = 4.76



d = 5.05





d = 6.12



d = 6.26



d = 6.26





d = 15.33



d = 15.57



d = 15.83

Proposed Metric: Coverage

Real

Fake Nearest Neighbours

Increasing distance





d = 4.78



d = 4.84



d = 4.91





d = 7.19



d = 7.31



d = 7.33





d = 13.74



d = 13.76



d = 13.98

Proposed Metric

Formulation

$$d(P_n, Q_m) = \mathbb{E}_{x \sim P_n} \left[\frac{1}{K} \sum_{j=1}^{K} \left\| f(x) - f(y_j^*) \right\|_2 \right]$$

Assume P_n to be samples from the generated distribution P and Q_m to be samples from the real distribution Q

1. $d(P_n, Q_m)$ measures how close generated samples are to real samples. We call this **Chamfer-Fidelity.**
2. $d(Q_m, P_n)$ measures how well the generated distribution covers the real distribution. We call this **Chamfer-Coverage.**

Lower is better. Two numbers; no isotropic balls, no yes/no test.

Thank You

Conclusion: Existing evaluation metrics for generative models are are either uninformative or unreliable.

We introduce a principled replacement that is intuitive and interpretable.

Website: serchirag.github.io/chamfer

