# Likelihood-based Out-of-Distribution Detection with Denoising Diffusion Probabilistic Models

**Joseph Goodier:** UKRI-CDT in Accountable, Responsible and Transparent AI (ART-AI)

**Professor Neill D. F, Campbell:** Department of Computer Science, Centre for the Analysis of Motion, Entertainment Research and Applications (CAMERA)

BMVC 2023

## 01 Abstract

Out-of-Distribution detection between dataset pairs has been extensively explored with generative models. We show that likelihood-based Out-of-Distribution detection can be extended to diffusion models by leveraging the fact that they, like other likelihood-based generative models, are dramatically affected by the input sample complexity. Currently, all Out-of-Distribution detection methods with Diffusion Models are reconstruction-based. We propose a new likelihood ratio for Out-of-Distribution detection with Deep Denoising Diffusion Models, which we call the Complexity Corrected Likelihood Ratio. Our likelihood ratio is constructed using Evidence Lower-Bound evaluations from an individual model at various noising levels. We present results that are comparable to state-of-the-art Out-of-Distribution detection methods with generative models.

## 02 Contributions

**In this paper, we present:**

- Evidence that input sample complexity dramatically affects the ELBO contributions from low noising levels in DDPMs, as is seen with other generative models.
- A likelihood-based OOD detection method using DDPMs. We use a likelihood ratio that is calculated using ELBO evaluations from low noise levels over the total ELBO from all noise levels. We define it to be the **(Complexity Corrected Likelihood Ratio)** (CCLR).

## 03 Complexity Correct Likelihood Ratio

Likelihood ratios have been applied extensively for likelihood-based Out-of-Distribution detection with generative models. Generative model likelihoods have been shown to be dramatically affected by the input complexity of a sample. This leads to higher likelihood scores for low complexity Out-of-Distribution samples and lower likelihood scores for high complexity In-Distribution samples. Likelihood ratios correct for this by removing the contribution from low-level image features that most contribute to image complexity. With generative models historically, this has been achieved by training a separate model that captures this information and then subtracting the second models contribution from the full model likelihood. We present a method that uses a single DDPM to construct a likelihood ratio, that we call CCLR:

$$CCLR_{k/T} = \mathcal{L}_{\theta}^{<k} - \mathcal{L}_{\theta}^{<T}$$

This leverages the fact that low-level image features emerge at low t-values and contribute an outsized amount to the likelihood estimates (See Fig. 1). The likelihood estimates from the low-level t-values, below some threshold k, are subtracted from the model likelihood.

## 04 Results

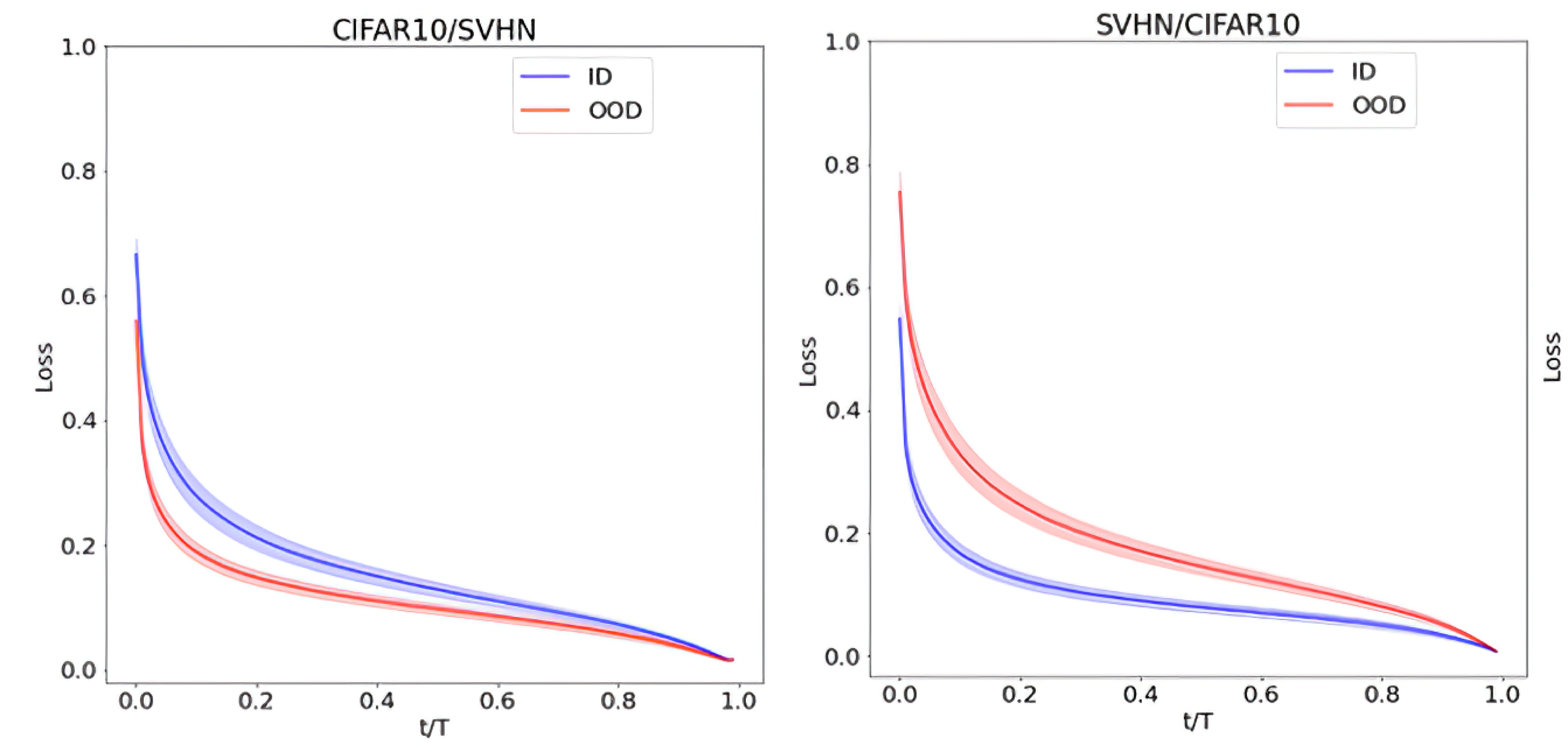


Figure 1: Plots showing the lower bound of the negative log-likelihood estimates sampled for a range of t-values averaged across ID and OOD test inputs. The higher complexity CIFAR10 has a larger negative log-likelihood **(lower likelihood)** when both ID and OOD class, than the less complex SVHN which displays a lower negative log-likelihood (**higher likelihood**).

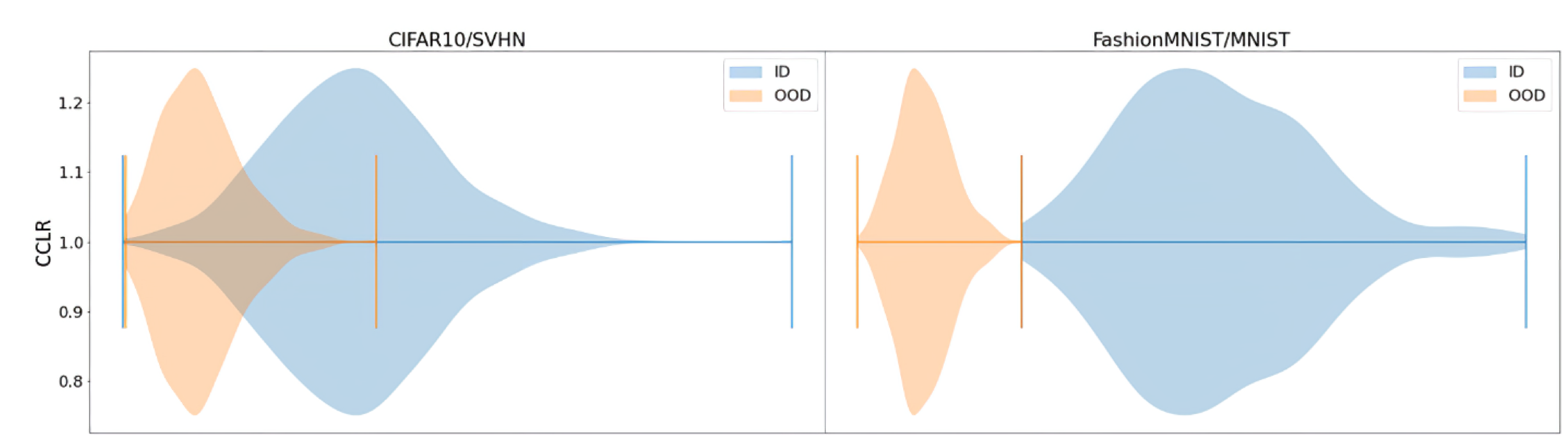


Figure 2: Violin plots displaying histograms of the CCLR score for both CIFAR10/SVHN (**Left**) and FashionMNIST/MNIST (**Right**) dataset pairs. For CIFAR10/SVHN violin plot, the CCLR scores were calculated using k/T = 1/2. For FashionMNIST/MNIST, CCLR scores were calculated using k/T = 1/5. Both of the selected k/T-values gave the highest AUROC scores for each respective dataset pair. These histograms relate to AUROC scores for each dataset pair separation of **0.964** for CIFAR10/SVHN (**Left**) and **1.00** for FashionMNIST/MNIST (**Right**).