# Out-of-Distribution Detection with a Single Unconditional Diffusion Model

*NeurIPS 2024*

**Alvin Heng, Alexandre H. Thiery, Harold Soh**

Collaborative, Learning, and Adaptive Robots

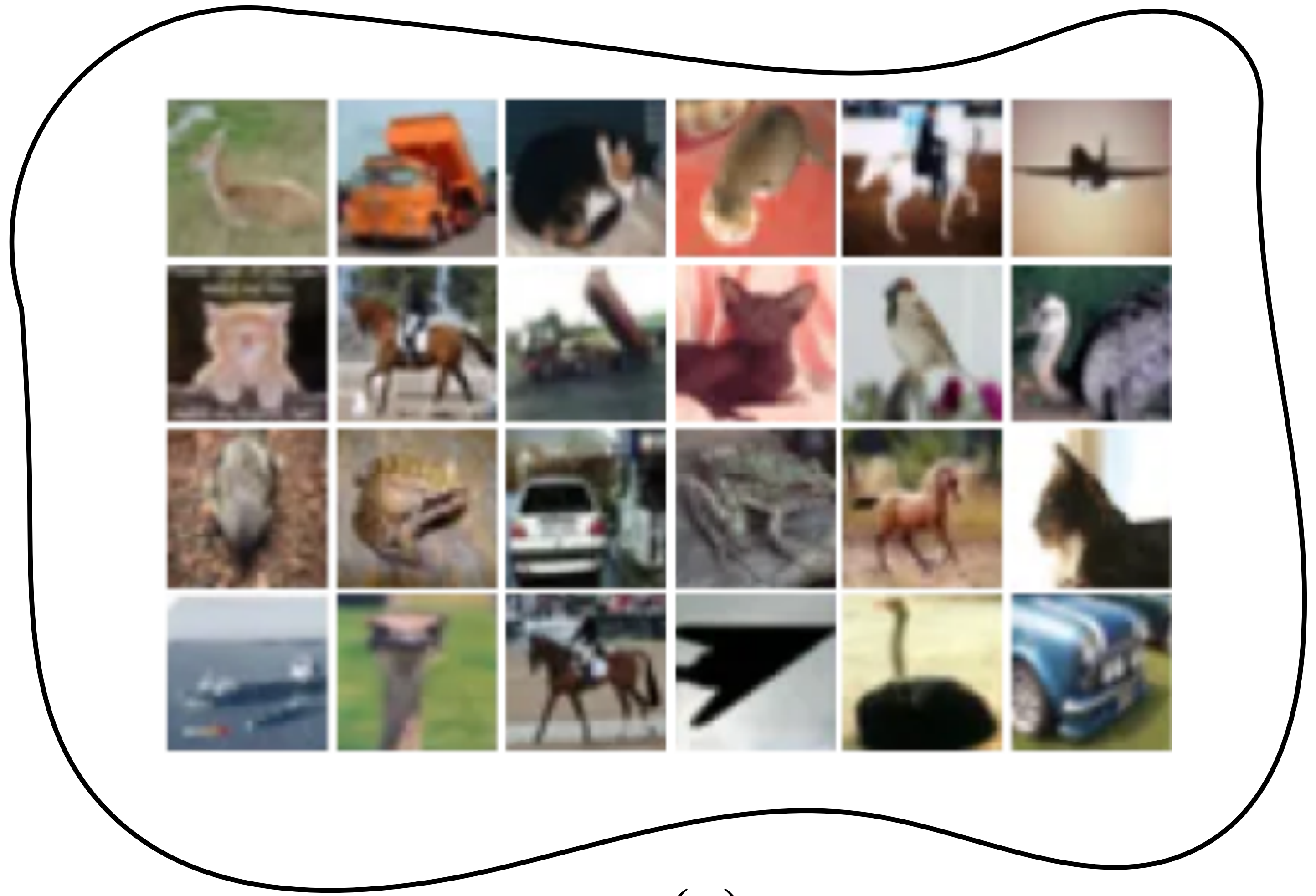
NUS National University of Singapore
School of Computing

NUS National University of Singapore
Department of Statistics and Data Science
Faculty of Science

# Out-of-Distribution Detection

- Given data samples $x_{train} \sim p(x)$, determine if a test sample $x_{test} \sim p(x)$.
- Deep neural networks shown to be overconfident on OOD samples.
- Train a generative model $p_\theta(x)$ from $x_{train}$ and evaluate $p_\theta(x_{test})$.

But likelihoods do not work!



$p(x)$


$x_{test}$

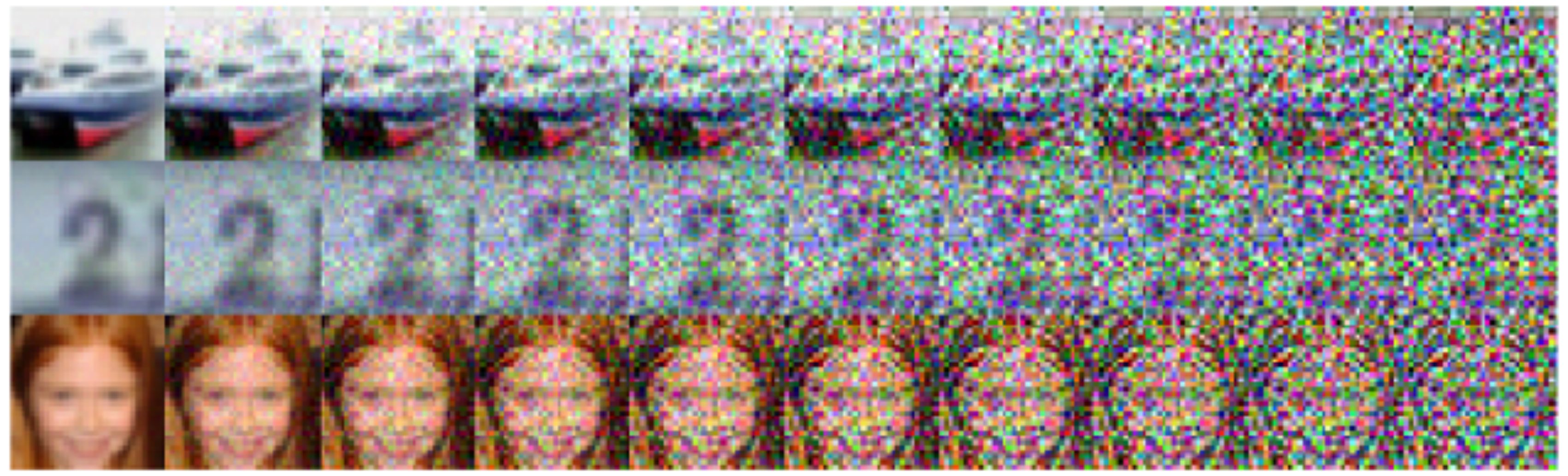
# Motivation: Scores for OOD Detection

- Assume two distributions $\phi_0(x)$ and $\psi_0(x)$ and their respective score estimates $\epsilon_\phi$ and $\epsilon_\psi$

$$D_{\mathrm{KL}}(\phi_0 \| \psi_0) = \frac{1}{2} \int_0^T \mathbb{E}_{\mathbf{x} \sim \phi_t} \frac{g(t)^2}{\sigma_t} \| \boldsymbol{\epsilon}_\phi(\mathbf{x}_t, t) - \boldsymbol{\epsilon}_\psi(\mathbf{x}_t, t) \|_2^2 \, \mathrm{d}t + D_{\mathrm{KL}}(\phi_T \| \psi_T).$$

- $\sum_t || \epsilon_{ID} - \epsilon_{OOD} ||^2$ different for different distributions => an OOD statistic, but we only have $\epsilon_{ID}$.

- Key insight: **a single model** can approximate scores for multiple distributions!



ImageNet Model

CelebA Model

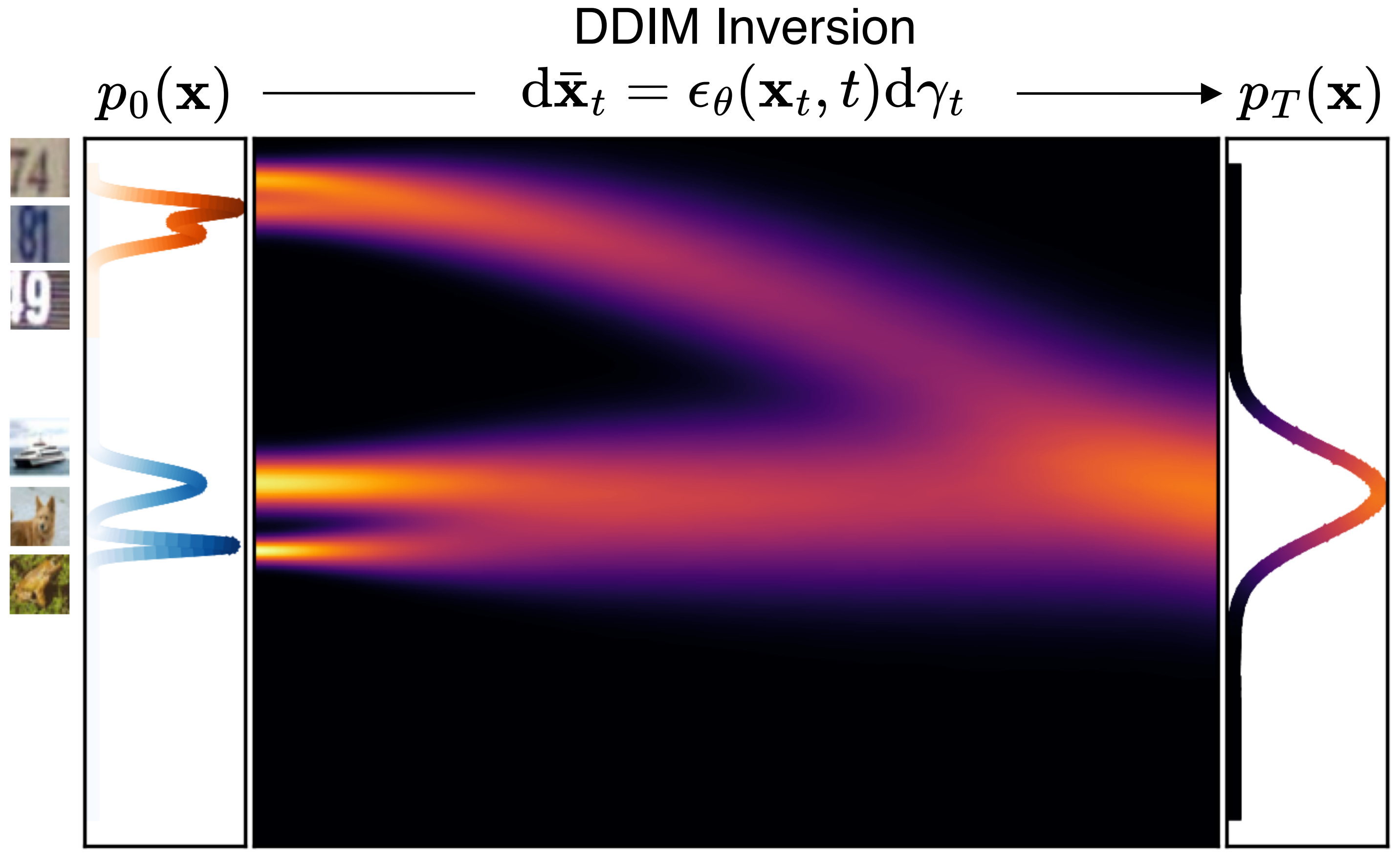
# DiffPath

$$\frac{\mathrm{d}\boldsymbol{\epsilon}_\theta}{\mathrm{d}\gamma_t} \approx \frac{\boldsymbol{\epsilon}_\theta(\mathbf{x}_{t+\Delta t}, t+\Delta t) - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)}{\Delta t}$$

- DDIM ODE:

$$\bar{\mathbf{x}}_{t_{n+1}} = \bar{\mathbf{x}}_{t_n} + h_n \boxed{\boldsymbol{\epsilon}_\theta(\mathbf{x}_{t_n}, t_n)} + \frac{1}{2!} h_n^2 \boxed{\frac{\mathrm{d}\boldsymbol{\epsilon}_\theta}{\mathrm{d}\gamma_t}\Big|_{(\bar{\mathbf{x}}_{t_n}, t_n)}} + \dots$$

- Measure the first ( $\sum_t ||\epsilon||^2$ ) and second ( $\sum_t ||d\epsilon/dt||^2$ ) derivatives of the diffusion path for OOD detection



# Pseudocode of DiffPath

**Algorithm 1** OOD detection with DiffPath

**Input:** Trained DM $\epsilon_\theta$, ID train set $\mathbf{X}_{\text{train}}$, test samples $\mathbf{X}_{\text{test}}$, empty lists $L_{\text{train}}$ and $L_{\text{test}}$
**Output:** OOD scores of test samples $S_\theta(\mathbf{X}_{\text{test}})$

1: **for** $\mathbf{x}_0$ in $\mathbf{X}_{\text{train}}$ **do**
2: $\{\epsilon_\theta(\mathbf{x}_t, t)\}_{t=0}^{T} \leftarrow \texttt{DDIMInversion}(\mathbf{x}_0, \epsilon_\theta)$ ▷ Integrate Eq. 7 from $t = 0$ to $T$
3: Calculate OOD statistic using $\{\epsilon_\theta(\mathbf{x}_t, t)\}_{t=0}^{T}$
4: Append statistic to $L_{\text{train}}$
5: **end for**
6: $p_{\text{train}}(\cdot) \leftarrow$ fit density estimate to $L_{\text{train}}$ ▷ e.g., KDE, GMM
7: $L_{\text{test}} \leftarrow$ Repeat lines $1 - 5$ with $\mathbf{X}_{\text{test}}$
8: **return** $p_{\text{train}}(l)$ for every $l$ in $L_{\text{test}}$

# Experimental Results

| Method | C10 *vs* | | | | SVHN *vs* | | | | CelebA *vs* | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SVHN | CelebA | C100 | Textures | C10 | CelebA | C100 | Textures | C10 | SVHN | C100 | Textures | Average | NFE |
| IC | 0.950 | 0.863 | 0.736 | - | - | - | - | - | - | - | - | - | - | - |
| IGEBM | 0.630 | 0.700 | 0.500 | 0.480 | - | - | - | - | - | - | - | - | - | - |
| VAEBM | 0.830 | 0.770 | 0.620 | - | - | - | - | - | - | - | - | - | - | - |
| Improved CD | 0.910 | - | **0.830** | 0.880 | - | - | - | - | - | - | - | - | - | - |
| DoS | 0.955 | 0.995 | 0.571 | - | 0.962 | **1.00** | 0.965 | - | 0.949 | 0.997 | 0.956 | - | 0.928 | - |
| WAIC[1] | 0.143 | 0.928 | 0.532 | - | 0.802 | 0.991 | 0.831 | - | 0.507 | 0.139 | 0.535 | - | 0.601 | - |
| TT[1] | 0.870 | 0.848 | 0.548 | - | 0.970 | **1.00** | 0.965 | - | 0.634 | 0.982 | 0.671 | - | 0.832 | - |
| LR[1] | 0.064 | 0.914 | 0.520 | - | 0.819 | 0.912 | 0.779 | - | 0.323 | 0.028 | 0.357 | - | 0.524 | - |
| *Diffusion-based* | | | | | | | | | | | | | | |
| NLL | 0.091 | 0.574 | 0.521 | 0.609 | **0.990** | 0.999 | **0.992** | 0.983 | 0.814 | 0.105 | 0.786 | 0.809 | 0.689 | 1000 |
| IC | 0.921 | 0.516 | 0.519 | 0.553 | 0.080 | 0.028 | 0.100 | 0.174 | 0.485 | 0.972 | 0.510 | 0.559 | 0.451 | 1000 |
| MSMA | 0.957 | **1.00** | 0.615 | **0.986** | 0.976 | 0.995 | 0.980 | **0.996** | 0.910 | 0.996 | 0.927 | **0.999** | **0.945** | **10** |
| DDPM-OOD | 0.390 | 0.659 | 0.536 | 0.598 | 0.951 | 0.986 | 0.945 | 0.910 | 0.795 | 0.636 | 0.778 | 0.773 | 0.746 | 350 |
| LMD | **0.992** | 0.557 | 0.604 | 0.667 | 0.919 | 0.890 | 0.881 | 0.914 | 0.989 | **1.00** | 0.979 | 0.972 | 0.865 | $10^4$ |
| *Ours* | | | | | | | | | | | | | | |
| DiffPath-6D-ImageNet | 0.856 | 0.502 | 0.580 | 0.841 | 0.943 | 0.964 | 0.954 | 0.969 | 0.807 | 0.981 | 0.843 | 0.964 | 0.850 | **10** |
| DiffPath-6D-CelebA | 0.910 | 0.897 | 0.590 | 0.923 | 0.939 | 0.979 | 0.953 | 0.981 | **0.998** | **1.00** | **0.998** | **0.999** | 0.931 | **10** |

# Summary

Propose to measure **rate-of-change** and **curvature** of diffusion paths for OOD detection.

Uses a **single model** across tasks as opposed to conventional methods requiring individually-trained models.

# For More Information

https://tinyurl.com/diffpath-paper

https://tinyurl.com/diffpath-code

alvin.heng@u.nus.edu or harold@comp.nus.edu.sg