Oops… There is a **distribution mismatch** in the **Street View House Numbers (SVHN)** dataset!

The SVHN Dataset Is Deceptive for Probabilistic Generative Models Due to a Distribution Mismatch

Tim Z. Xiao[1,2,*] Johannes Zenn[1,2,*] Robert Bamler[1]
[1]University of Tübingen [2]IMPRS-IS [*]Equal contribution, order determined by coin flip.

Table 1 : For SVHN, we find that the FID between random subsets of the training and test set is **significantly higher** than the FID between non-overlapping subsets of the training set of the same size, while the IS for $\mathcal{D}'_{\text{train}}$ and $\mathcal{D}'_{\text{test}}$ is similar within all datasets.

FID (↓), IS (↑)	SVHN	SVHN-Remix	CIFAR-10
$\text{FID}(\mathcal{D}''_{\text{train}}, \mathcal{D}'_{\text{train}})$	3.309 ± 0.029	3.334 ± 0.018	5.196 ± 0.040
$\text{FID}(\mathcal{D}''_{\text{train}}, \mathcal{D}'_{\text{test}})$	$\mathbf{16.687 \pm 0.325}$	3.326 ± 0.015	5.206 ± 0.031
$\text{IS}(\mathcal{D}'_{\text{train}} \mid \bar{\mathcal{D}}_{\text{train}})$	8.507 ± 0.114	8.348 ± 0.568	7.700 ± 0.043
$\text{IS}(\mathcal{D}'_{\text{test}} \mid \bar{\mathcal{D}}_{\text{train}})$	8.142 ± 0.501	8.269 ± 0.549	7.692 ± 0.023

Defining Distribution Mismatch

- **Assumption:** $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ consist of i.i.d. samples from an underlying distribution $p_{\text{data}}(x)$.
- With a *distance metric* D, we expect $D\big(p_{\text{data}}(x), \mathcal{D}'_{\text{train}}\big) \approx D\big(p_{\text{data}}(x), \mathcal{D}'_{\text{test}}\big)$, where $\mathcal{D}'_{\text{train}}$ and $\mathcal{D}'_{\text{test}}$ are equally sized random subsets of $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$.
- **Evaluation:** We use $\mathcal{D}''_{\text{train}}$ representing $p_{\text{data}}(x)$ and compute whether $D\big(\mathcal{D}''_{\text{train}}, \mathcal{D}'_{\text{train}}\big) \approx D\big(\mathcal{D}'_{\text{train}}, \mathcal{D}'_{\text{test}}\big)$.
- $\text{FID}(\mathcal{D}_1, \mathcal{D}_2) = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2^2 + \text{Tr}\left(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2 - 2\left(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2\right)^{1/2}\right)$
- $\text{IS}(\mathcal{D} \mid \mathcal{D}_{\text{train}}) = \exp\left(\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\big[D_{\text{KL}}\left[p_{\text{cls.}}(y \mid \boldsymbol{x}) \parallel p_{\text{cls.}}(y)\right]\big]\right)$

Summary

- There is a distribution mismatch in SVHN! (I.e., training and test set do not come from the same distribution.)
- The distribution mismatch affects the evaluation of probabilistic generative models, but not classifiers.
- **Lesson:** When benchmarking generative models, we need to be mindful of distribution mismatch!
- We provide the **SVHN-Remix** dataset.

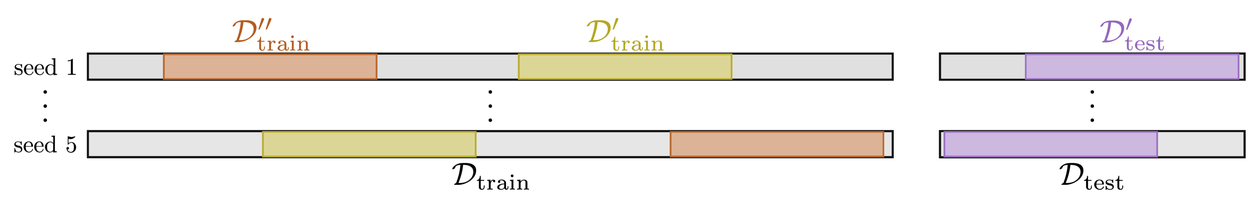
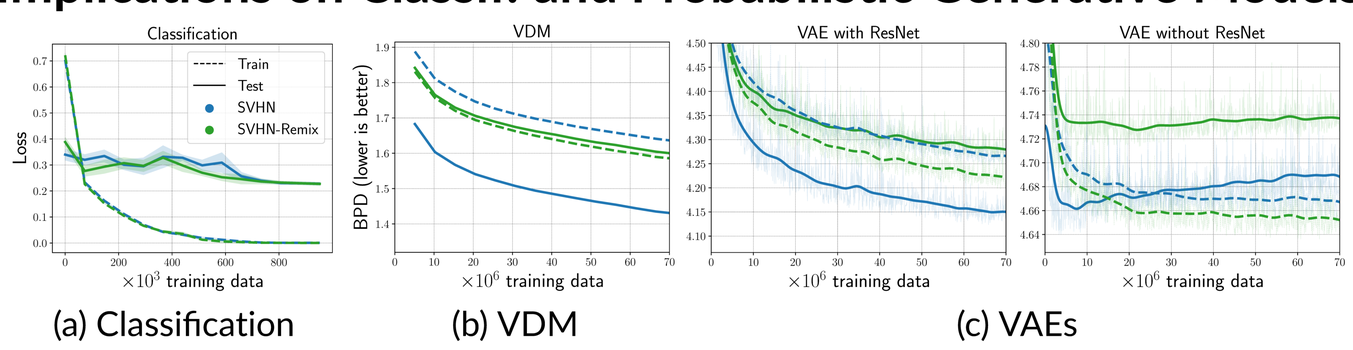


Figure 1 : Five random splits (with reshuffling) of $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ into $\mathcal{D}'_{\text{train}}$, $\mathcal{D}''_{\text{train}}$, and $\mathcal{D}'_{\text{test}}$.

Implications on Classif. and Probabilistic Generative Models



(a) Classification (b) VDM (c) VAEs

Figure 2 : (a): classification loss evaluated on training set (dashed) and test set (solid) on SVHN and SVHN-Remix (shaded areas are $\pm\sigma$). The losses are similar.
(b) and (c): BPD evaluated as a function of training progress on the training set and test set for a variational diffusion model (VDM) and variational autoencoders (VAEs). For SVHN, the order of training and test set performance is flipped compared to SVHN-Remix.

- Bits per dimension (BPD; proportional to negative ELBO; lower is better).
- The solid blue line first goes below the dashed blue line, then goes above it ⇒ overfitting!

SCAN ME

Scan QR code to checkout the project website!

Presentation at NeurIPS 2023 Workshop on Distribution Shifts (DistShift)

EBERHARD KARLS UNIVERSITÄT TÜBINGEN