Multimodal base distributions in conditional flow matching generative models

Shane Josias & Willie Brink
Applied Mathematics, Stellenbosch University, South Africa
josias@sun.ac.za, wbrink@sun.ac.za

Introduction

Background. Continuous normalising flows specify a target distribution $p_x(\boldsymbol{x})$ in terms of an easy-to-sample-from base distribution $p_u(\boldsymbol{u})$, an invertible transformation given by the solution to a neural ordinary differential equation:

$$\frac{d\boldsymbol{z}(t)}{dt} = \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t), \quad t \in [t_1, t_0], \quad \boldsymbol{z}_1 = \boldsymbol{z}(t_1) = \boldsymbol{x}, \quad \boldsymbol{z}_0 = \boldsymbol{z}(t_0) = \boldsymbol{u}, \tag{1}$$

and the instantaneous change of variables formula [1]:

$$\frac{\partial}{\partial t} \log p(\boldsymbol{z}) = -\mathrm{Tr}\left(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{z}}\right). \tag{2}$$

We show that continuous normalising flows, trained through the conditional flow matching objective (CFM models), provide unreliably high likelihoods on out-of-distribution data (Figure 1).

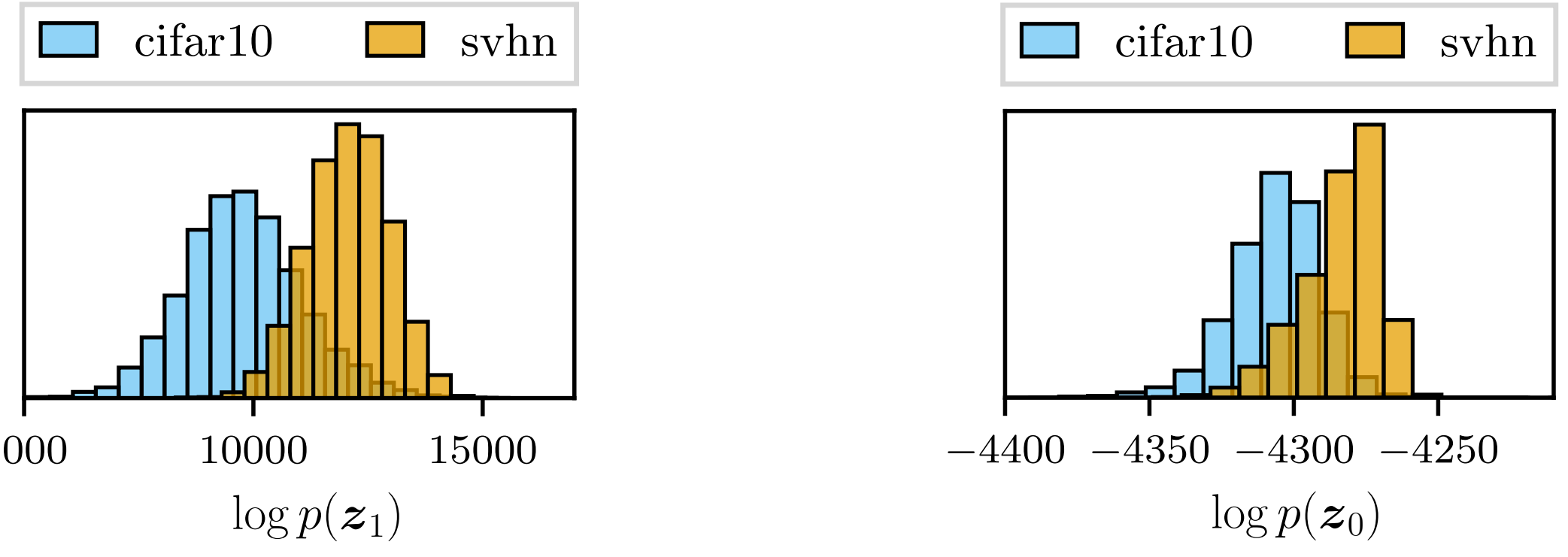


Figure 1: Out-of-distribution samples from SVHN are assigned higher likelihoods $(p(\boldsymbol{z}_1))$ compared to in-distribution samples from CIFAR10. We see these undesirably high likelihoods also under the model's unimodal base distribution $(p(\boldsymbol{z}_0))$, prompting our investigation into multimodal base distributions.

Research question. We investigate whether a multimodal Gaussian mixture model (GMM) base distribution can lead to more reliable out-of-distribution likelihoods. As motivation, we show in Figure 2 that a CFM model might easily transform out-of-distribution data points between separate modes in the target space to a region of high-likelihood under a unimodal base distribution. With a class-informed multimodal base distribution, we hope that the model can assign appropriately low likelihood to out-of-distribution data.

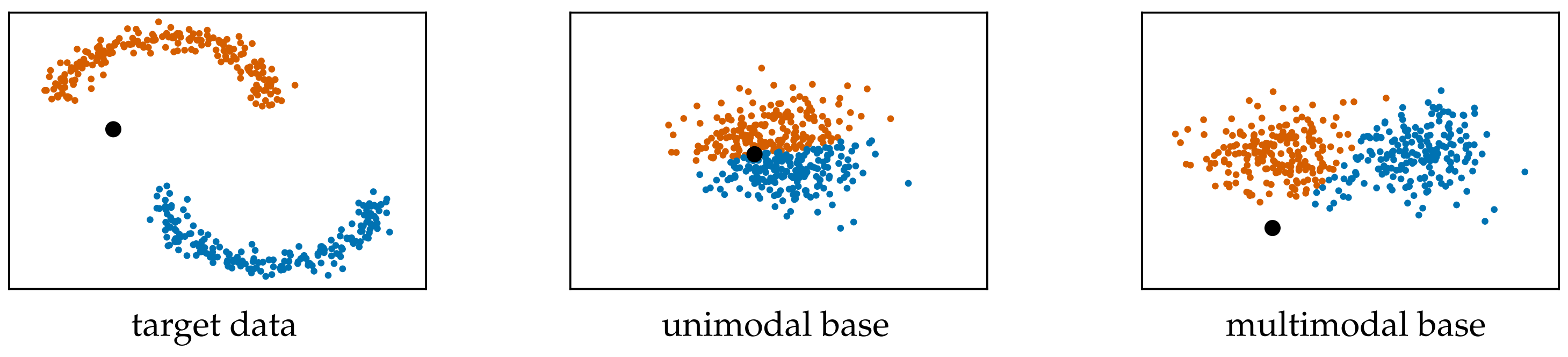

target data
unimodal base
multimodal base

Figure 2: For CFM models trained on a 2D "moons" dataset (left), we see an out-of-distribution test point (black dot) being transformed to a point with high likelihood under the standard unimodal base distribution (middle), and low likelihood under a multimodal base distribution (right).

Contributions. The GMM base enables sampling from the target distribution in a class-specific manner, performs comparable to a standard (unimodal) Gaussian base, but may not be sufficient to solve the problem of high likelihoods for out-of-distribution data. We also find that CFM models may depend too strongly on pixel values, rather than semantic content.

Methodology

Given a class-labelled training set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ with $\boldsymbol{x}_i \in \mathbb{R}^d$, and letting $\boldsymbol{z}_1 = \boldsymbol{x}_i$, we construct a continuous flow that computes the log-likelihood of a test sample by simultaneously solving Equations 1 and 2 for $t \in [t_1, t_0]$. The flow models are trained through the conditional flow matching objective [2]

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} ||\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t) - \boldsymbol{u}_t(\boldsymbol{z}_t \mid \boldsymbol{z}_1)||^2, \tag{3}$$

with Gaussian probability paths defined over $t \sim \mathcal{U}(0, 1)$. The dynamics function $f_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t)$ is implemented as a U-Net with attention.

Standard base distribution. A standard base can be implemented through the following target conditional vector field, with σ_{min} sufficiently small:

$$\boldsymbol{u}_t(\boldsymbol{z}_t \mid \boldsymbol{z}_1) = \frac{\boldsymbol{z}_1 - (1 - \sigma_{\text{min}})\boldsymbol{z}_t}{1 - (1 - \sigma_{\text{min}})t}. \tag{4}$$

GMM base distribution. To incorporate multimodality and class information, we consider a GMM base distribution with a component for each of the K classes in the data:

$$p_u(\boldsymbol{z}_0) = \sum_{k=1}^{K} c_k \mathcal{N}(\boldsymbol{z}_1 \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{5}$$

with $\boldsymbol{\mu}_k$ set to the empirical mean of each class represented in the training set, $\boldsymbol{\Sigma}_k = \sigma^2\boldsymbol{I}$, and c_k the relative class frequencies. This GMM base can be implemented through the following target conditional vector field:

$$\boldsymbol{u}_t(\boldsymbol{z}_t \mid \boldsymbol{z}_1) = \frac{\boldsymbol{z}_1 - \sigma_{\text{min}}\boldsymbol{\mu}_k - (1 - \sigma_{\text{min}})\boldsymbol{z}_t}{1 - (1 - \sigma_{\text{min}})t}. \tag{6}$$

Log likelihoods

Bits-per-dimension scores in Table 1 shows that the GMM base distribution performs comparably on in-distribution data, but does not alleviate the problem of unreliable likelihoods. We show through a shuffling experiment in Figure 3 that CFM models may depend too strongly on pixel values.

Table 1: Bits-per-dimension scores for CFM models trained on various datasets, for in- and out-of-distribution test sets, when using the standard (unimodal) and GMM base distributions. A lower bpd implies a higher likelihood on the data under the model.

CFMs trained on MNIST

	Standard	GMM
MNIST-Test	1.15 ± 0.01	1.73 ± 0.04
FashionMNIST-Test	4.68 ± 0.02	5.13 ± 0.15

CFMs trained on CIFAR10

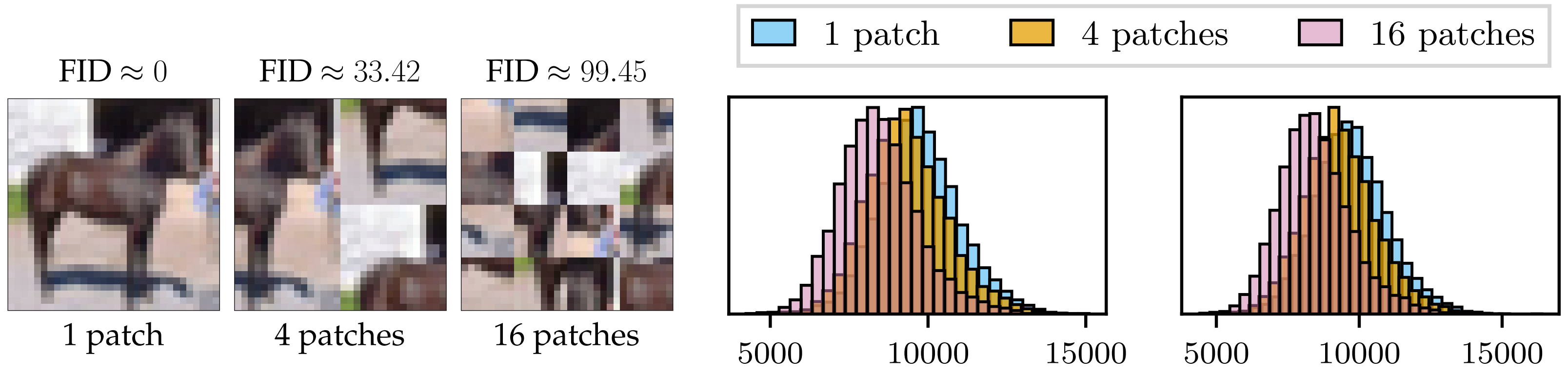
	Standard	GMM
CIFAR10-Test	3.42 ± 0.01	3.50 ± 0.01
SVHN-Test	2.32 ± 0.01	2.41 ± 0.01

CFMs trained on FashionMNIST

	Standard	GMM
FashionMNIST-Test	2.87 ± 0.01	3.39 ± 0.06
MNIST-Test	1.75 ± 0.02	2.29 ± 0.06

CFMs trained on SVHN

	Standard	GMM
SVHN-Test	2.11 ± 0.00	2.20 ± 0.01
CIFAR10-Test	3.83 ± 0.01	3.94 ± 0.01



(a) shuffling of an example image

(b) $\log p(\boldsymbol{z}_1)$, with standard (left) and GMM (right) bases

Figure 3: In-distribution test images are randomly shuffled, as illustrated in (a), leading to the histograms of log-likelihoods shown in (b) under models that use the standard (left) and GMM (right) base distributions. Test log likelihood histograms are less affected than FID scores for out-of-distribution patch-shuffled datasets, indicating an over-reliance on pixel values.

Sample quality

Table 2 shows FID scores for generated samples. Across all the datasets, models using the standard base produce samples of higher quality compared to the GMM base (with a large outlier for the SVHN model that uses the standard base). The high FID for the GMM can be attributed to mode-collapse, and is surprising given the class-specific parameterisation of the base distribution. Figure 4 shows a few generated samples from the various models.

Table 2: Fréchet inception distances (FID) of generated samples from CFM models trained with the standard and GMM base distributions. Lower is better. We include in the last column results from models that use a GMM base with larger covariance scaling, leading to a base distribution that is approximately unimodal.

Dataset	Standard	GMM	GMM ($\sigma^2 = 1$)
MNIST	3.20 ± 1.25	20.18 ± 7.31	2.00
FashionMNIST	5.37 ± 0.83	57.04 ± 6.64	7.50
CIFAR10	27.62 ± 1.78	64.85 ± 12.52	29.66
SVHN	33.10 ± 39.85	62.20 ± 21.54	50.06

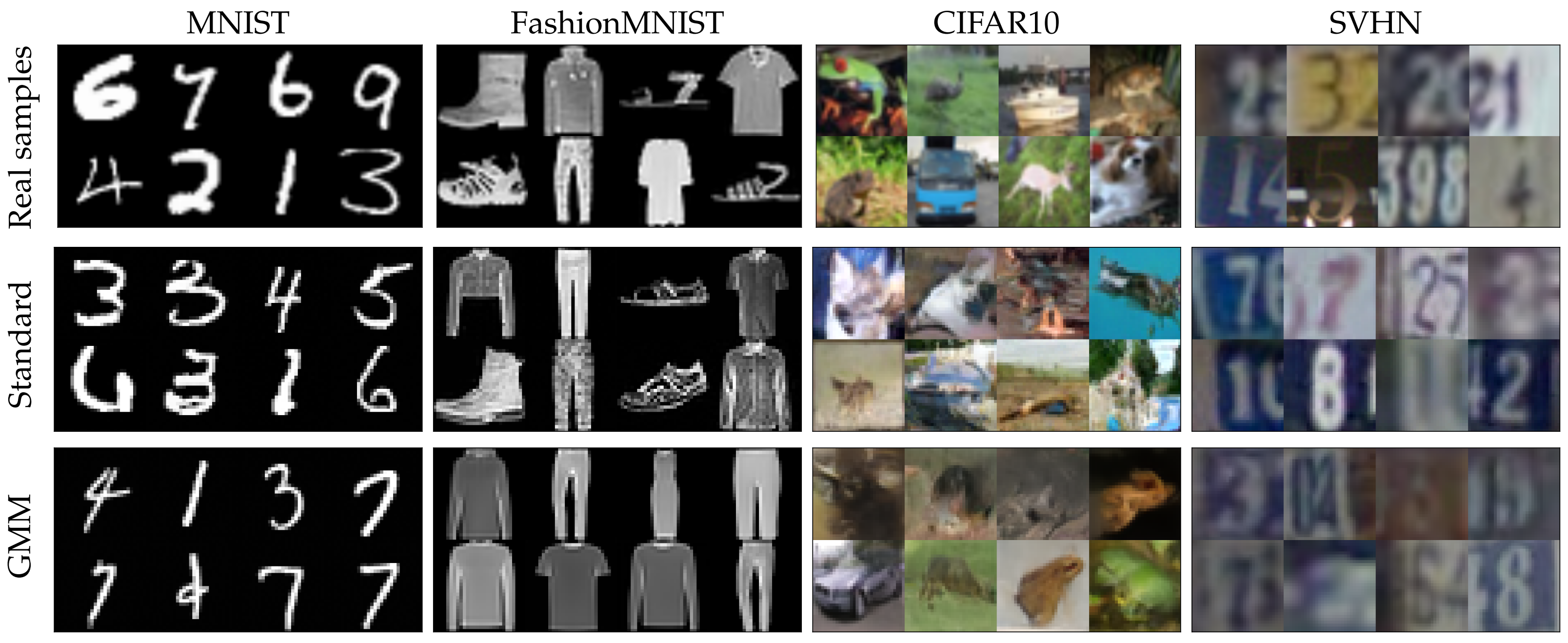


Figure 4: Real and generated samples from the best performing CFM models (according to FID scores over 50K samples) with the standard and GMM base distributions.

Future work

Future work could include training CFM models in a feature space with semantic consistency, to circumvent their dependence on pixel frequencies. It may also be possible to learn the GMM parameters, to mitigate the observed mode-collapse in generated samples.

Acknowledgement

This work is based on research supported by the National Research Foundation of South Africa (grant number 138341).

References

[1] Ricky T.Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 2018.

[2] Yaron Lipman, Ricky T.Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. *International Conference on Learning Representations*, 2023.