Multimodal base distributions in conditional flow matching generative models

Shane Josias
josias@sun.ac.za
Willie Brink
wbrink@sun.ac.za

Department of Mathematical Sciences
Stellenbosch University
Stellenbosch, South Africa

Abstract

Normalising flows are a flexible class of generative models that provide exact likelihoods, and are often trained through maximum likelihood estimation. Recent work suggests that these models can assign undesirably high likelihood to out-of-distribution data, questioning their reliability for applications where likelihoods are important (e.g. outlier detection). We show that continuous normalising flows trained with the conditional flow matching objective, instead of maximum likelihood, also provide unreliable likelihoods. We then argue for and investigate the utility of incorporating multimodality in the base distribution, through a Gaussian mixture model (GMM) centred at the empirical means of a target distribution's modes. The GMM has an additional benefit in that samples can be generated from specified modes. We find that the GMM base distribution leads to performance comparable to a standard (unimodal) base distribution for in- and out-of-distribution likelihoods, at little to no extra cost in training and inference times. Interestingly, samples generated by models that use a GMM base have higher precision but significantly lower recall compared to the standard base. We also find support for the hypothesis that continuous flows depend too strongly on pixel values, rather than semantic content.

1 Introduction

Normalising flows are generative models that specify a target density through a base distribution and an invertible transformation process, and have been successfully applied for image [7, 14, 19] and video [17] generation, computer graphics [22] and sensor noise modelling [1]. They offer exact likelihood evaluation as an advantage over other generative models, enabling, in principle, the ability for outlier detection. We are specifically interested in the reported phenomenon of normalising flows assigning undesirably high likelihoods to out-of-distribution data [15, 24, 32], which questions their reliability in applications.

A discrete-step normalising flow specifies a target distribution $p_x(\boldsymbol{x})$ in terms of an easy-to-sample-from base distribution $p_u(\boldsymbol{u})$, and an invertible transformation $\boldsymbol{u} = g(\boldsymbol{x})$ with $\boldsymbol{u} \sim p_u(\boldsymbol{u})$, by employing the change-of-variables formula. $g(\boldsymbol{x})$ is defined as a composite function, usually chosen to be a neural network whose architecture is restricted for a tractable log-determinant in the change-of-variables formula. The continuous-time variant [4, 9] (hereafter referred to as a continuous flow) expresses $\boldsymbol{u} = g(\boldsymbol{x})$ as the solution

to an initial value problem (IVP):

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

and uses a continuous analog of the change of variables formula to determine $\log p_x(\boldsymbol{z}_1)$ [4]. The function $f_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t)$ defines a time-dependent vector field describing the transformation dynamics, with trainable parameters $\boldsymbol{\theta}$. This formulation circumvents restrictions on $g(\boldsymbol{x})$ for a tractable log-determinant, at the time-cost of simulating solution trajectories for the IVP in Equation 1. For a chosen base distribution and transformation function, the flow is trained by maximum likelihood. It is sufficient for the base distribution $p_u(\boldsymbol{z}_0)$ to be a standard unimodal Gaussian [4, 6, 7, 8, 9, 14, 16, 26, 32], however, when trained with the maximum likelihood objective, these models can provide unreliable likelihoods for out-of-distribution data [15, 24, 32]. For instance, a model trained on CIFAR10 may provide higher likelihoods for samples from SVHN (similar to what we show in Figure 1).

Continuous flows trained with the recently introduced conditional flow matching [19] objective (CFM models) circumvent maximum likelihood training and the need for simulating solution trajectories. With the bottleneck of simulation removed, continuous flows become more relevant to applications at scale. We show for the first time that CFM models also provide unreliable out-of-distribution likelihoods; an undesirable phenomenon in view of the benefit that flows provide in terms of exact likelihood evaluation. Then, motivated by the considerable overlap in the base distribution likelihoods for both in- and out-of-distribution data, as seen in Figure 1 for example, we investigate whether a multimodal base distribution can lead to more reliable out-of-distribution likelihoods. As further illustration, 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.

We incorporate multimodality through a Gaussian mixture model (GMM), with component means centred at the empirical means of the target distribution's modes. Our work complements existing approaches by reporting both in- and out-of-distribution likelihoods on common image datasets, with a view towards understanding the out-of-distribution failure modes of CFM models. We also include sample quality and diversity metrics to evaluate whether multimodality in the base distribution is beneficial for data generation.

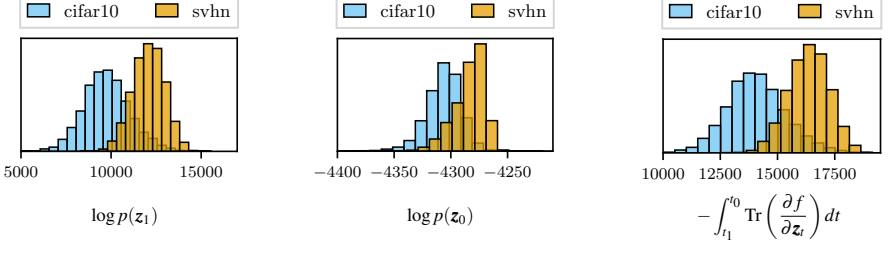


Figure 1: Decomposed log-likelihood histograms for data under a continuous-time flow model trained on the CIFAR10 dataset, with $\log p(\boldsymbol{z}_1) = \log p(\boldsymbol{z}_0) - \int_{t_1}^{t_0} \mathrm{Tr}\left(\partial f / \partial \boldsymbol{z}_t\right) dt$. Out-of-distribution samples from the SVHN dataset are assigned higher likelihoods $(p(\boldsymbol{z}_1))$ compared to in-distribution samples. We see these undesirably high likelihoods also under the model's standard base distribution $(p(\boldsymbol{z}_0))$, prompting our investigation into an alternative.

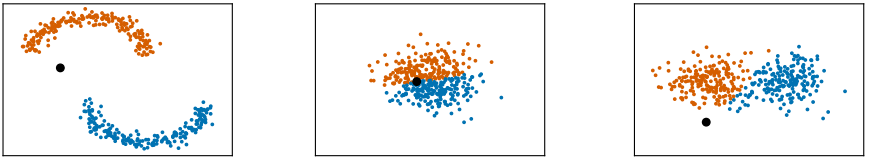


Figure 2: For CFM models trained on a 2D "moons" dataset (left), we see an out-of-distribution test point (black dot) being transformed to a point with high likelihood under the standard unimodal base distribution (middle), and low likelihood under a multimodal base distribution (right). Orange and blue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

Even though a GMM base enables sampling from the target distribution in a class-specific manner, our results suggest that it may not be sufficient to solve the problem of high likelihoods for out-of-distribution data, and indeed performs comparable to a standard (unimodal) Gaussian base. We also find that CFM models may depend too strongly on pixel values, rather than semantic content, suggesting interesting avenues for future work.

2 Related work

That normalising flows provide unreliable likelihoods for out-of-distribution samples has been noted before. The seminal work of Nalisnick et al. [24] showed that discrete-step flows such as Glow [14] and RealNVP [7] assign higher likelihoods to out-of-distribution data, according to the bits-per-dimension metric. Kirichenko et al. [15] further investigated this behaviour, specifically for RealNVP [7], and found that the inductive biases of the flow can influence out-of-distribution likelihoods. The same phenomenon has been observed for continuous flows [12, 32]. These models are all trained through maximum likelihood, making it hard to determine whether the training objective is the cause of unreliable likelihoods. Our work complements the literature on out-of-distribution likelihoods by showing that CFM models (trained with the flow matching objective rather than through maximum likelihood) exhibit similar behaviour. It is worth highlighting the importance of this observation, given that CFM models are more scalable than simulation-based continuous flows. Our results also suggest that maximum likelihood training might not be the cause of unreliable likelihoods.

We depart from the recent literature on conditional flow matching by focusing on the base distribution, instead of the probability path parameterisation [2, 20, 27, 31]. As far as we know, this is the first work that considers multimodality in the base distribution for CFM models. Our work adapts the Gaussian probability paths introduced by Lipman et al. [19] to incorporate multimodality in the form of a GMM, with a view towards understanding the failure modes of out-of-distribution likelihoods for continuous flows. GMMs have been used for discrete-step flows for the task of density estimation and semi-supervised image classification [15, 25, 29], and have yielded improvements for those tasks. It is also common to use class labels to improve image synthesis in other generative models, like generative adversarial networks (GANs) and diffusion models. For GANs, label information can be included via a projection in the discriminator [21] or self-attention in the generator [33]. For diffusion models, it can be included via adaptive group normalisation, or through other class conditional processes [5].

3 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 as

$$\log p(\boldsymbol{z}_1) = \log p(\boldsymbol{z}_0) - \int_{t_1}^{t_0} \mathrm{Tr}\left[\frac{\partial f}{\partial \boldsymbol{z}_t}\right] dt. \tag{2}$$

Following Grathwohl et al. [9], the transformed sample $\boldsymbol{u} = \boldsymbol{z}_0$ and $\log p(\boldsymbol{z}_1)$ are obtained by simultaneously solving Equations 1 and 2 for $t \in [t_1, t_0]$. Hutchinson's trace approximation is applied to the Jacobian term for computational efficiency. Equation 2 describes how probability paths $p_t(\boldsymbol{z}_t)$ between the target and base densities evolve over time. We restrict our focus to Gaussian conditional probability paths,

$$p_t(\boldsymbol{z}_t \mid \boldsymbol{z}_1) = \mathcal{N}(\boldsymbol{z}_t \mid \boldsymbol{\mu}_1(t), \sigma^2(t)\boldsymbol{I}), \tag{3}$$

where $\boldsymbol{\mu}_1(t)$ and $\sigma^2(t)$ describe how the mean and covariance change over time, with $\boldsymbol{\mu}_1$ parameterised by $\boldsymbol{z}_1$. Such a probability path follows trajectories between a density concentrated around $\boldsymbol{z}_1$ and the base density, and is specified by a conditional vector field (Lipman et al. [19], Theorem 3):

$$\boldsymbol{u}_t(\boldsymbol{z}_t \mid \boldsymbol{z}_1) = \frac{\sigma'(t)}{\sigma(t)}(\boldsymbol{z}_t - \boldsymbol{\mu}_1(t)) + \boldsymbol{\mu}_1'(t), \tag{4}$$

where the prime symbol indicates the derivative with respect to t. The conditional flow matching [19] objective is then defined as

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} ||f_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t) - \boldsymbol{u}_t(\boldsymbol{z}_t \mid \boldsymbol{z}_1)||^2, \tag{5}$$

with 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 [11, 19]. The target conditional vector field must be set so that there is a valid probability path between the data distribution $p_1(\boldsymbol{z}_1)$ and the base distribution $p_0(\boldsymbol{z}_0)$.

Standard base distribution. As a baseline, we consider the standard unimodal Gaussian base distribution. For the Gaussian probability paths in Equation 3, a standard base can be constructed by defining $\boldsymbol{\mu}_1(t) = t\boldsymbol{z}_1$ and $\sigma(t) = 1 - (1 - \sigma_{\min})t$, leading to the following target conditional vector field:

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

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:

$$p_u(\boldsymbol{z}_0) = \sum_{k=1}^{K} c_k \mathcal{N}(\boldsymbol{z}_1 \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{7}$$

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. Probability paths that lead to a component mean in the

GMM can be constructed by defining $\boldsymbol{\mu}_1(t) = t\boldsymbol{z}_1 + (1-t)\boldsymbol{\mu}_k$ and $\sigma(t) = 1-(1-\sigma_{\min})t$, with $\boldsymbol{\mu}_k$ referring to the class mean of sample $\boldsymbol{z}_1$. This leads to the following target conditional vector field (we provide a derivation in the supplementary material):

$$\boldsymbol{u}_t(\boldsymbol{z}_t \mid \boldsymbol{z}_1) = \frac{\boldsymbol{z}_1 - \sigma_{\min}\boldsymbol{\mu}_k - (1-\sigma_{\min})\boldsymbol{z}_t}{1-(1-\sigma_{\min})t}. \tag{8}$$

Equation 8 defines a probability path between a density centred at the data point $\boldsymbol{z}_1$, and a Gaussian centred at the empirical mean of the class associated with $\boldsymbol{z}_1$. $\sigma_{\min}$ is set sufficiently small so that the density is concentrated around a sample. In practice, samples are obtained from the GMM component corresponding to each class in a batch during training. At test time, the likelihood of a sample is evaluated and weighed across all components, as in Equation 7. We use the log-sum-exp trick for numerical stability.

Datasets. We consider the MNIST, FashionMNIST, CIFAR10, and SVHN datasets, for which it has been shown that flow models provide unreliable likelihoods [24]. Refer to Figure 5 (top row) for samples from each dataset.

Likelihood metric. Bits-per-dimension (bpd) is used to evaluate in- and out-of-distribution likelihoods, computed as

$$\text{bpd} = -\frac{\log_2 p(\boldsymbol{x})}{d} = -\frac{\log p(\boldsymbol{x})}{d\log(2)}, \tag{9}$$

where d is the dimension of the data and $\log p(\boldsymbol{x})$ is averaged over a test set. This metric gives an indication of the number of bits required, on average, to encode the data under the model [23]. A higher bpd implies a lower average likelihood under the model. Histograms of per-sample log-likelihoods over the test set will also be considered, as in Figure 1.

Sample quality. The Fréchet inception distance (FID) [10] is used as a measure of sample quality and diversity, computed between the training set and 50K generated samples, in accordance with common practice [3, 13, 18]. A lower FID implies better sample quality, but a high FID can mean either that the model is not able to generate high quality samples, or that it captures only a subset of the data modes [23]. Therefore we also consider precision and recall, as proposed by Kynkäänniemi et al. [18]: precision measures whether generated samples are as close (or closer) in feature space to the training samples as training samples are to one another, and recall measures whether training samples are as close (or closer) in feature space to generated samples as generated samples are to one another.

Quantitative metrics are calculated over multiple runs, with models trained over 150 epochs. We use hyperparameters from Lipman et al. [19] as a starting point, reduce the capacity of the models slightly due to compute constraints, and perform hyperparameter tuning on the learning rate and batch size for both base distributions. In addition, we perform hyperparameter tuning on the GMM covariance scale σ^2, restricting it to values between 0.5 and 0.8. We observed that decreasing the covariance for the GMM too much can lead to poorer samples, and $\sigma^2 \geq 1$ leads to a GMM base that is approximately the unimodal standard base for data scaled in the interval $(0,1)$. Code to reproduce our results is available at this https URL.

4 Results

Table 1 shows bits-per-dimension scores for separate models trained on the four datasets. Our results reproduce for CFM models what has been observed for discrete flows [15, 24] and continuous flows trained through maximum likelihood [32]. CFM models with standard and GMM base distributions trained on FashionMNIST, assign higher likelihoods (lower bpd) to MNIST data, and models trained on CIFAR10 assign higher likelihoods to SVHN data. The fact that these models are trained with a flow matching objective suggests that maximum likelihood training might not be the cause of unreliable likelihoods reported previously [32]. We further observe that CFM models trained on MNIST or SVHN provide reliable likelihoods, again corroborating what has been shown for discrete [15, 24] and continuous flows trained through maximum likelihood [32].

Table 1: Bits-per-dimension scores for conditional flow matching (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. Means and standard deviations are measured over multiple training runs.

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

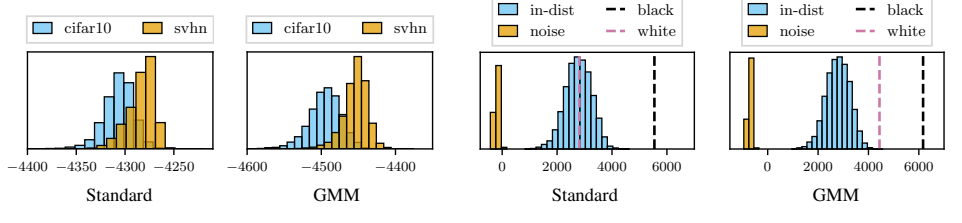
CFMs trained on FashionMNIST

	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

CFM models trained with a GMM base distribution provide likelihoods comparable to what is achieved with the standard base distribution, at no additional computational cost during training or inference. We do recognise that the standard base distribution works surprisingly well on multi-class image data, despite not incorporating any class information, and that the use of a multimodal base distribution does not alleviate the problem of unreliable



(a) $\log p(\boldsymbol{z}_0)$ from CFMs trained on CIFAR10 (b) $\log p(\boldsymbol{z}_1)$ from CFMs trained on FashionMNIST

Figure 3: Histograms of test data log-likelihoods under the base distribution (a) and under the target distribution (b), for the training sets and base distributions as indicated.

likelihoods for out-of-distribution data.

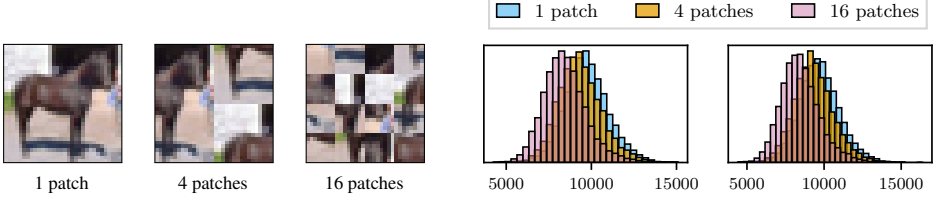
Our interest in multimodal base distributions was initially motivated by a significant overlap of likelihoods assigned to in- and out-of-distribution data under a model that uses a unimodal base. However, as we see in Figure 3(a), using a GMM does not necessarily improve matters. Likelihoods assigned to out-of-distribution data remain higher compared to in-distribution data, giving further merit to the suggestion that the problem may be due to the data distribution [24] or the complexity of the data itself [28, 32]. We explore this hypothesis for FashionMNIST, where the majority of training image pixels are close in intensity to black (0) or white (255). In Figure 3(b) we show how CFM models trained on FashionMNIST respond to an image of constant intensity 0 ("black"), an image of constant intensity 255 ("white"), and images whose pixel intensities are uniformly random between 0 and 255 ("noise"). A model with the GMM base assigns higher likelihoods to constant images compared to the standard base. Centering the base distribution modes on the empirical class means may encourage the model to assign higher likelihoods to more frequently occurring pixels, rather than the semantic content. Interestingly, with the standard base, white images are mapped to the most common likelihood values of the FashionMNIST test set. As a sanity check, we see that images with random intensities have the lowest likelihoods.

Following Voleti et al. [32], we also test the degree to which conditional flow matching models rely on pixel values, rather than semantic content, by inspecting likelihoods on datasets of shuffled patches. Figure 4(a) shows an example test image from the CIFAR10 dataset, alongside shuffled versions that use 4 patches and 16 patches respectively. If FID_p denotes the Fréchet inception distance score for a version of the CIFAR10 test set shuffled using p patches, we find

$$\text{FID}_1 \approx 0, \qquad \text{FID}_4 = 33.42, \qquad \text{FID}_{16} = 99.45.$$

This shows that shuffling images by 4 patches already influences the FID score significantly, compared to the unshuffled baseline FID_1. Care is taken to ensure that unshuffled images are not in the dataset of shuffled patches.

The log-likelihood histograms in Figure 4(b) show that a CFM model with either the standard base or the GMM base assigns marginally lower likelihoods for the shuffled test images. The overlap with likelihoods of the unshuffled (1 patch) test data is striking, considering how much the FID is affected by this type of shuffling. Although sample quality and likelihoods are independent [30], a large change in FID for shuffled images may indicate that the semantic content of the images are significantly changed. We might hope that the likelihoods



(a) shuffling of an example image (b) $\log p(\boldsymbol{z}_1)$, with standard (left) and GMM (right) bases

Figure 4: 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.

respond in a similar way, and that it does not may indicate an over-dependence on pixel content, as also seen for the FashionMNIST dataset in Figure 3(b). Such an over-dependence is now further motivated by the observation that likelihood histograms are invariant under shuffling of image patches, for both the standard and the GMM base distributions.

For an indication of generated sample quality, we show in Table 2 the FID for 50K generated samples, from models trained on each of the four datasets, using the standard and GMM bases. 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). We note that it is possible to obtain comparable FID scores for models that use the GMM base, by setting the covariance scaling in $\mathbf{\Sigma}_k$ (Equation 7) to $\sigma^2 = 1$. However, given the overall scale of our base distributions, where points have coordinates in the interval $(0, 1)$, this value of σ^2 would yield a GMM that is approximately unimodal. It is curious that generated samples with the GMM base distribution do appear visually similar to real samples, as demonstrated in Figure 5, despite large FID scores. We also remark that sample quality can potentially be increased by training for longer.

Table 2: Fréchet inception distances of generated samples from CFM models trained with the standard and GMM base distributions. We include in the last column results from models that use a GMM base with larger covariance scaling.

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

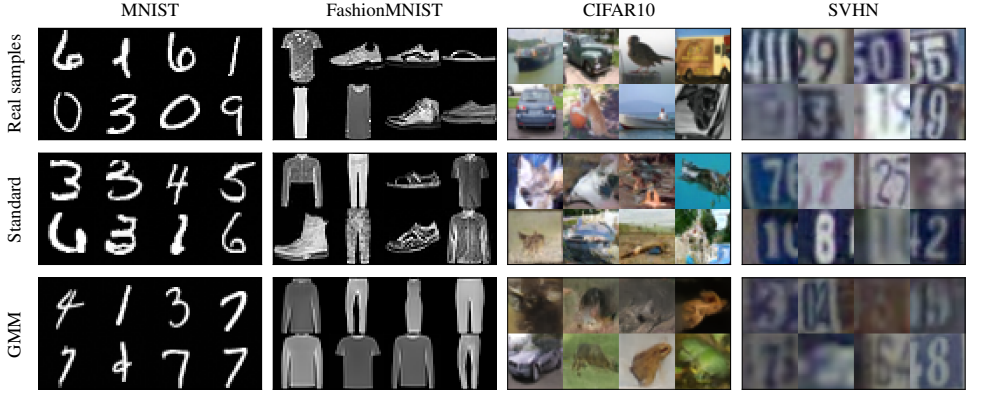


Figure 5: Real and generated samples from the best performing CFM models (according to FID scores over 50K samples) with the standard and GMM base distributions.

Inspecting high FIDs. As mentioned earlier, there can be different factors influencing high FIDs. Because of this, we report in Table 3 the precision and recall of generated samples from a few specific models. Precision measures the realism of generated samples, while

recall measures the degree to which the modes of the training data are covered [13]. For CIFAR10, we select the best performing model for each base distribution, to determine why models with a GMM base distribution lead to a high FID. For SVHN, we select the worst performing models to compare the two base distributions (for this dataset, and the standard base, one of the models in our training runs led to a large outlier in FID). For completeness, we also include models that use the GMM with large covariance. The results in Table 3 suggest that, in fact, the GMM base leads to CIFAR10 samples that are more realistic (higher precision) compared to the standard base. The higher FID scores can therefore be attributed to mode-collapse, given the very low recall, which is surprising since we set the GMM component means to the empirical means of the classes. For SVHN, the outlying FID score can be attributed to a low precision. The GMM base consistently leads to mode collapse, but produces more realistic samples, except for when we set the covariance scale to 1.

Table 3: Precision and recall scores for samples generated by our CFM models. Higher is better, to a maximum of 1. For CIFAR10 we show results from the best performing models over multiple training runs, and for SVHN the worst. The last row contains scores for samples from random training runs (neither the best nor the worst), with the covariance scaling for the GMM base set to 1.

Dataset	Precision		Recall	
	Standard	GMM	Standard	GMM
CIFAR10 (best)	0.55	0.78	0.24	0.04
SVHN (worst)	0.32	0.64	0.50	0.23
SVHN (random; $\sigma^2 = 1$ for GMM)	0.47	0.38	0.49	0.49

5 Conclusion and future work

Driven by an observation that high likelihoods for out-of-distribution data persist in the base distribution of continuous-time normalising flow models trained with the conditional flow matching objective, we explored whether multimodality in the base distribution may be beneficial. Our results indicate that a multimodal base performs comparably to the standard (unimodal) base, and may not be sufficient to alleviate the problem of unreliable out-of-distribution likelihoods. CFM models with multimodal base distributions generate more realistic samples but, surprisingly, suffer from mode collapse. This provides avenues for further research. It might be worth investigating whether other (possibly learned) parameterisations of our GMM base distribution might allow for a better precision/recall trade-off. Indeed, class-labelled data simplifies the parameterisation of the GMM, but it is still an open question as to whether unlabelled data can be handled in a sensible way.

We also showed that CFM models may depend too strongly on pixel values, rather than semantic content. It could be more effective to instead apply CFM models to a latent space with semantic consistency. Kirichenko et al. [15] showed that it is possible to circumvent unreliable likelihoods when training discrete flow models on features from a pre-trained classifier, at the cost of the ability to generate samples. We hypothesise a similar result will hold for CFM models, and aim to verify this in future for a latent space that maintains the ability to generate samples.

In conclusion, our work contributes to the narrative on reliable likelihoods from scalable continuous flow models.

Acknowledgement

This work is based on research supported by the South African Department of Higher Education and Training, and the National Research Foundation (grant number 138341).

References

[1] Abdelrahman Abdelhamed, Marcus A. Brubaker, and Michael S. Brown. Noise flow: Noise modeling with conditional normalizing flows. *International Conference on Computer Vision*, 2019.

[2] Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *International Conference on Learning Representations*, 2023.

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. *International Conference on Learning Representations*, 2019.

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

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 2021.

[6] Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: Non-linear independent components estimation. *International Conference on Learning Representations, Workshop Track*, 2015.

[7] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using RealNVP. *International Conference on Learning Representations*, 2017.

[8] Chris Finlay, Jörn-Henrik Jacobsen, Levon Nurbekyan, and Adam M. Oberman. How to train your neural ODE: The world of Jacobian and kinetic regularization. *International Conference on Machine Learning*, 2020.

[9] Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. FFJORD: Free-form continuous dynamics for scalable reversible generative models. *International Conference on Learning Representations*, 2019.

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*, 2017.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020.

[12] Shane Josias and Willie Brink. Multimodal base distributions for continuous-time normalising flows. *Advances in Neural Information Processing Systems, The Symbiosis of Deep Learning and Differential Equations Workshop*, 2023.

[13] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[14] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in Neural Information Processing Systems*, 2018.

[15] Polina Kirichenko, Pavel Izmailov, and Andrew G. Wilson. Why normalizing flows fail to detect out-of-distribution data. *Advances in Neural Information Processing Systems*, 2020.

[16] Ivan Kobyzev, Simon J.D. Prince, and Marcus A. Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979, 2020.

[17] Manoj Kumar, Mohammad Babaeizadeh, Dumitru Erhan, Chelsea Finn, Sergey Levine, Laurent Dinh, and Durk Kingma. Videoflow: A conditional flow-based model for stochastic video generation. *International Conference on Learning Representations*, 2020.

[18] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 2019.

[19] 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.

[20] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *International Conference on Learning Representations*, 2023.

[21] Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. *International Conference on Learning Representations*, 2018.

[22] Thomas Müller, Brian McWilliams, Fabrice Rousselle, Markus Gross, and Jan Novák. Neural importance sampling. *ACM Transactions on Graphics*, 38(5):1–19, 2019.

[23] Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023.

[24] Eric T. Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Görür, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? *International Conference on Learning Representations*, 2019.

[25] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. *Advances in Neural Information Processing Systems*, 2017.

[26] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(1):2617–2680, 2021.

[27] Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky T.Q. Chen. Multisample flow matching: Straightening flows with minibatch couplings. *International Conference on Machine Learning*, 2023.

[28] Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F. Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. *International Conference on Learning Representations*, 2020.

[29] Vincent Stimper, Bernhard Schölkopf, and José Miguel Hernández-Lobato. Resampling base distributions of normalizing flows. *International Conference on Artificial Intelligence and Statistics*, 2022.

[30] Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. *International Conference on Learning Representations*, 2016.

[31] Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024.

[32] Vikram Voleti, Chris Finlay, Adam Oberman, and Christopher Pal. Multi-resolution continuous normalizing flows. *Annals of Mathematics and Artificial Intelligence*, 2024.

[33] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *International Conference on Machine Learning*, 2019.