"Best of Many" Samples Distribution Matching

Apratim Bhattacharyya, Mario Fritz, Bernt Schiele
Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken, Germany
{abhattac, mfritz, schiele}@mpi-inf.mpg.de

Abstract

Generative Adversarial Networks (GANs) suffer from the mode collapse problem. Variational Autoencoders explicitly maximize a reconstruction-based data log-likelihood forcing the coverage of all modes, but at the cost of the overall quality of generated samples. Recent efforts to combine VAE and GAN frameworks still suffer from several issues chief among them that the data log-likelihood and prior terms are at odds. To deal with this, we propose a novel objective for VAE-GAN frameworks which integrates a "Best-of-Many" sample reconstruction cost. Our proposed objective along with our hybrid VAE-GAN framework shows significant improvement over both prior hybrid VAE-GANs and plain GANs in mode coverage and quality.

1 Introduction

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) have achieved state-of-the-art performance, with respect to sample quality e.g. in case of generative modeling of image distributions. However, GANs do not explicitly estimate the data likelihood and thus is no direct incentive to cover the whole data distribution leading to the mode collapse problem.

In contrast, Variational Autoencoders (VAEs) (Kingma and Welling, 2014) explicitly maximize data likelihood and can be forced to cover all modes (Bozkurt et al., 2018; Shu et al., 2018). VAEs enable sampling by constraining the latent space to a unit Gaussian prior. However, the L_1/L_2 reconstruction based data likelihood leads to lower overall sample quality – blurriness in case of images. There has been a spur of recent work Donahue et al. (2017); Larsen et al. (2016); Rosca et al. (2019) which aims integrate GANs in a VAE framework to improve VAE generation quality. Notably in Rosca et al. (2019), GANs are integrated in a VAE framework by augmenting the L_1/L_2 data likelihood term in the VAE objective with a GAN discriminator based synthetic likelihood ratio term.

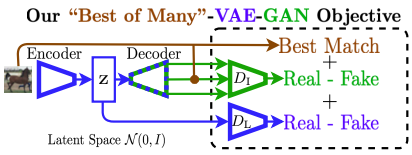


Figure 1: In contrast to prior work, our novel objective gives multiple chances to the encoder to draw samples with high likelihood.

However, Rosca et al. (2019) report that in case of hybrid VAE-GANs, the latent space does not usually match the Gaussian prior. This is because, the reconstruction log-likelihood in the VAE objective is at odds with the divergence to the latent prior (Tabor et al., 2018) (also in case of alternatives proposed by Makhzani et al. (2016); Arjovsky et al. (2017)). This problem is further exacerbated with the addition of the synthetic likelihood term in the hybrid VAE-GAN objective – it is necessary for sample quality but it introduces additional constraints on the encoder/decoder. This leads to the degradation in the quality and diversity of samples. Therefore, we relax the constraints on the hybrid VAE-GAN encoder, giving the encoder multiple chances to draw samples with high reconstruction likelihood – only the best sample is penalized (see Figure 1). We show that our "Best of Many"-VAE-GAN framework significantly improves upon prior hybrid VAE-GANs and plain GANs, on the benchmark highly multi-modal synthetic datasets, CIFAR-10 and CelebA.

4th workshop on Bayesian Deep Learning (NeurIPS 2019), Vancouver, Canada.

2 Leveraging the "Best of Many" samples

We begin with a discussion of our VAE backbone with prior distribution $p(\mathrm{z})$. VAEs maximize the log-likelihood of the data, using amortized variational inference with a recognition network, where $q_\phi(\mathrm{z}|\mathrm{x})$ is the posterior distribution of latent variables. The ELBO is maximized,

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})} \log(p_\theta(\mathrm{x}|\mathrm{z})) - \mathrm{D}_{KL}(q_\phi(\mathrm{z}|\mathrm{x}) \parallel p(\mathrm{z})). \tag{1}$$

However, in (1) the average likelihood of the samples generated from the posterior $q_\phi(\mathrm{z}|\mathrm{x})$ is maximized. This forces all samples from $q_\phi(\mathrm{z}|\mathrm{x})$ to explain x equally well, penalizing any variance in $q_\phi(\mathrm{z}|\mathrm{x})$ and thus forcing it away from the Gaussian prior $p(\mathrm{z})$. Therefore, in Bhattacharyya et al. (2018) a "Best of Many" sample objective was proposed,

$$\mathcal{L}_{\text{MS}} = \log \left(\mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})} p_\theta(\mathrm{x}|\mathrm{z}) \right) - \mathrm{D}_{KL}(q_\phi(\mathrm{z}|\mathrm{x}) \parallel p(\mathrm{z})). \tag{2}$$

Unlike (1), this objective considers multiple samples from $q(\mathrm{z}|\mathrm{x})$ and the likelihood of only the best sample. This allows $q_\phi(\mathrm{z}|\mathrm{x})$ to have higher variance, helping it better match the prior and significantly reducing the trade-off with the data log-likelihood. However, only a reconstruction based Gaussian data likelihood term is considered, which might not be sufficient in case of complex high dimensional distributions e.g. in case of image data this leads to blurry samples. We address this issue by integrating a GAN based synthetic likelihood term into (2), (full derivation in the Appendix)

$$\begin{aligned} \mathcal{L}_{\text{MS}} &\propto \alpha \log \left(\mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})} \frac{p_\theta(\mathrm{x}|\mathrm{z})}{p(\mathrm{x})} \right) + \beta \log \left(\mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})} p_\theta(\mathrm{x}|\mathrm{z}) \right) - \mathrm{D}_{KL}(q_\phi(\mathrm{z}|\mathrm{x}) \parallel p(\mathrm{z})) \\ &\approx \alpha \log \left(\mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})} \frac{p_\theta(\mathrm{y}=1|\mathrm{z},\mathrm{x})}{1 - p(\mathrm{y}=1|\mathrm{x})} \right) + \beta \log \left(\mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})} p_\theta(\mathrm{x}|\mathrm{z}) \right) - \mathrm{D}_{KL}(q_\phi(\mathrm{z}|\mathrm{x}) \parallel p(\mathrm{z})). \end{aligned} \tag{3}$$

The ratio $p_\theta(\mathrm{y}=1|\mathrm{z},\mathrm{x}) / 1 - p(\mathrm{y}=1|\mathrm{x})$ should be high for generated samples which are indistinguishable from real images and low otherwise. This can be realized using a shared jointly trained discriminator D_{I}. In case of image data, we observe that directly estimating this ratio using D_{I} leads to increased stability and improved results and can be further improved with the addition of spectral normalization (Miyato et al., 2018). In practice, we estimate both the integrals of (3) using Monte-Carlo integration. The reconstruction based likelihood $\mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})} p_\theta(\mathrm{x}|\mathrm{z})$ takes the form $e^{-\lambda \|\mathrm{x} - \hat{\mathrm{x}}\|_n}$ – a log-sum-exp which is numerically unstable. As we perform stochastic gradient descent, we can deal with this after sampling of the data points. We can well estimate the log-sum-exp using the max – the "Best-of-Many" samples (Nielsen and Sun, 2016). The final objective optimized by our BMS-VAE-GAN takes the form,

$$\mathcal{L}_{\text{BMS-S}} = \alpha \ \mathbb{E}_i \log \left(D_{\mathrm{I}}(\mathrm{x}|\hat{\mathrm{z}}^i) \right) + \beta \max_i \log(p_\theta(\mathrm{x}|\hat{\mathrm{z}}^i)) - \mathrm{D}_{KL}(q_\phi(\mathrm{z}|\mathrm{x}) \parallel p(\mathrm{z})). \tag{4}$$

We use the same optimization scheme as in Rosca et al. (2019). We provide the algorithm in detail in the Appendix.

Approximation error. This "Best-of-Many" estimation does introduce a $\log(T)$ error term. However, as pointed out in Bhattacharyya et al. (2018), this error term is dominated by the low data likelihood term in the beginning of optimization. Later, as generated samples become more diverse, the log likelihood term is dominated by the best sample – "Best of Many" is equivalent.

Classifier based estimate of the prior term. Recent work (Makhzani et al., 2016; Arjovsky et al., 2017; Rosca et al., 2019) has shown that point-wise minimization of the KL-divergence using its analytical form leads to degradation in image quality. Instead, KL-divergence term is recast in a synthetic likelihood ratio form (similar to (3)) minimized "globally" using a classifier instead of point-wise. Therefore, unlike Bhattacharyya et al. (2018), here we employ a classifier based estimate of the KL-divergence to the prior. However, as pointed out by prior work on hybrid VAE-GANs (Rosca et al., 2019), a classifier based estimate with still leads to mismatch to the prior as the trade-off with the data log-likelihood still persists without the use of the "Best of Many" samples. Therefore, as we shall demonstrate next, the benefits of using the "Best of Many" samples extends to case when a classifier based estimate of the KL-divergence is employed.

3 Experiments

Next, we evaluate our novel objective (4) for hybrid VAE-GANs on highly multi-modal synthetic data, CIFAR-10 and CelebA.

Table 1: Evaluation on multi-modal synthetic data.

Method	2D Grid (25 modes)		2D Ring (8 modes)	
	Modes	HQ%	Modes	HQ%
VEEGAN (Srivastava et al., 2017)	24.6	40	8	52.9
GDPP-GAN (Elfeki et al., 2019)	24.8	68.5	8	71.7
SN-GAN (Miyato et al., 2018)	23.8±1.5	90.9±4.0	6.8±1.1	86.6±9.7
MD-GAN (Eghbal-zadeh et al., 2019)	25	99.3±2.2	8	89.0±3.6
WAE (Arjovsky et al., 2017)	25	65.4±3.8	8	35.8±4.7
α-GAN (Rosca et al., 2019)	25	70.5±4.2	8	83.6±5.3
BMS-VAE-GAN (Ours) $T = 10$	25	**99.7±0.2**	8	**99.6±0.3**

Table 2: Visualization of samples.

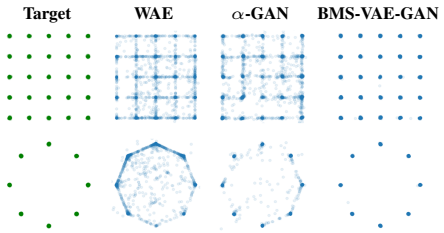
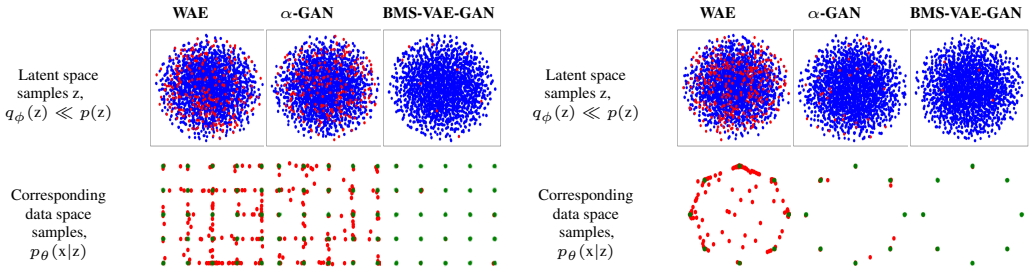


Table 3: Effect of our novel objective in the latent space. **Top Row:** The standard WAE and α-GAN objectives leads to mismatch to the prior in the latent space. We show samples z (in red) which are highly likely under the standard Gaussian prior (blue) but have low probability under the learnt marginal posterior $q_\phi(\mathrm{z})$. **Bottom Row:** We show that such points z lead to low quality data samples (in red), which do correspond to any of the modes.



Synthetic data. We evaluate in Table 1 on the standard 2D Grid and Ring datasets, which are highly challenging due to their multi-modality. The metrics considered are the number of modes captured and % of high quality samples (within 3 standard deviations of a mode). The generator/discriminator architecture is same as in Srivastava et al. (2017). We see that our BMS-VAE-GAN (using the best of $T = 10$ samples) outperforms state of the art GANs e.g. (Eghbal-zadeh et al., 2019) and the WAE and α-GAN baselines. The explicit maximization of the data log-likelihood enables our BMS-VAE-GAN and the WAE and α-GAN baselines to capture all modes in both the grid and ring datasets. The significantly increased proportion of high quality samples with respect to WAE and α-GAN baselines is due to our novel "Best of Many" objective. We illustrate this in Table 3. Following Rosca et al. (2019) we analyze the learnt latent spaces in detail, in particular we check for points (in red) which are likely under the Gaussian prior $p(\mathrm{z})$ (blue) but have low probability under the marginal posterior $q_\phi(\mathrm{z}) = \int q_\phi(\mathrm{z}|\mathrm{x})dx$. We use tSNE to project points from our 32-dimensional latent space to 2D. In Table 3 (Top Row) we clearly see that there are many such points in case of the WAE and α-GAN baselines (note that this low probability threshold is common across all methods). In Table 3 (Bottom Row) we see that these points lead to the generation of low quality samples (in red) in the data space. Therefore, we see that our "Best of Many" samples objective helps us match the prior in the latent space and thus this leads to the generation of high quality samples and outperforming both state of the art GANs and hybrid VAE-GAN baselines.

Table 4: FID on CIFAR-10 using 10k/5k real/generated samples.

Method	FID ↓
100k Generator Iterations	
SN-GAN (Miyato et al., 2018)	25.5
BW-GAN (Adler and Lunz, 2018)	25.1
α-GAN + SN (Rosca et al., 2019) $T = 1$	24.6±0.3
BMS-VAE-GAN (Ours) $T = 10$	23.8±0.2
BMS-VAE-GAN (Ours) $T = 30$	**23.4±0.2**
300k Generator Iterations	
Dist-GAN (Tran et al., 2018)	22.9
BMS-VAE-GAN (Ours) $T = 10$	**21.8±0.2**

Table 5: FID on CelebA using 10k/10k real/generated samples.

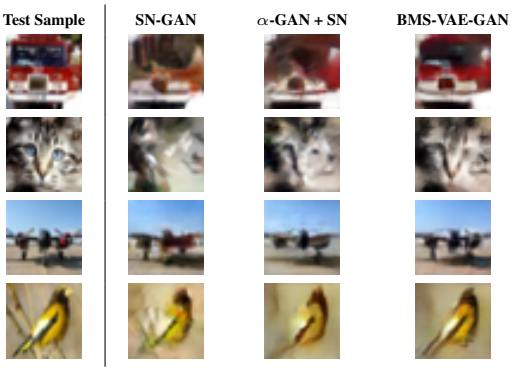
Method	FID ↓
DCGAN (Radford et al., 2016)	31.1±0.9
WGAN-GP (Gulrajani et al., 2017)	26.8±1.2
BEGAN (Berthelot et al., 2017)	26.3±0.9
Dist-GAN (Tran et al., 2018)	23.7±0.3
SN-GAN (Miyato et al., 2018)	21.9±0.8
α-GAN (Rosca et al., 2019)	19.2±0.8
α-GAN + SN (Rosca et al., 2019) $T = 1$	15.9±0.2
BMS-VAE-GAN (Ours) $T = 30$	**14.7±0.4**

CIFAR-10. The popular CIFAR-10 dataset consists of images belonging to 10 diverse real-world classes with considerable intra-class variability. It has been observed that plain auto-encoding based approaches Kingma and Welling (2014); Makhzani et al. (2016) do not perform well, as a simple Gaussian reconstruction based likelihood is insufficient for such highly multi-modal image data. We report quantitative results in Table 4, Table 6 using the FID and IoVM metrics. Please note that due to the much higher dimensionality (100), it makes the latent spaces much harder to reliably analyze. Therefore, we rely on the FID and IoVM metrics – also as in this task we are primarily interested in sample (image) quality. We use the standard CNN architecture used in SN-GAN (Miyato et al., 2018) and the hinge loss to optimize the synthetic likelihood. Therefore, we compare to models which also use the same model architecture and the hinge loss on the adversarial discriminator. We better highlight the effectiveness of our "Best of Many" objective, we compare to a improved version of α-GAN (compared to the simple DCGAN based model used in (Rosca et al., 2019)) with a standard CNN architecture and hinge loss on the discriminator. We see that we outperform both the hybrid α-GAN and plain SN-GAN and BW-GAN in terms of FID score. This shows that our BMS-VAE-GAN is better at capturing the image distribution without any loss of sample quality. We also observe an increase in performance with $T = 30$ samples, although this increase saturates with increasing T. The IoVM metric (also see Table 7) again illustrates that we are able to better capture the image distribution. Finally, we also experiment using the setting of Dist-GAN Tran et al. (2018) – training for 300k iterations with the same generator architecture. Again, we significantly outperform Dist-GAN and achieve a FID of 21.8 – the state of the art FID score with hinge loss.

Table 6: IvOM on CIFAR-10.

Method	IvOM ↓
DCGAN (Radford et al., 2016)	0.0084±0.0020
VEEGAN (Srivastava et al., 2017)	0.0068±0.0001
SN-GAN (Miyato et al., 2018)	0.0055±0.0006
α-GAN + SN (Rosca et al., 2019) $T = 1$	0.0048±0.0005
BMS-VAE-GAN (Ours) $T = 30$	**0.0037±0.0005**

Table 7: Closest generated images found using IvOM.



CelebA. We present quantitative results on the CelebA dataset in Table 5 and qualitative results in Appendix C. We again employ the FID metric to measure performance. We use the standard DCGAN architecture across all baselines. We use the hinge loss to optimize the synthetic likelihood. We again compare to a improved version of α-GAN with a hinge loss on the discriminator to highlight the effectiveness of the "Best of Many" samples. We again observe that our BMS-VAE-GAN with $T = 30$ samples perform best. We see that the base DCGAN has the weakest performance among the GANs. BEGAN suffers from partial mode collapse. The SN-GAN performs better compared to the WGAN-GP. This shows the effectiveness of Spectral Normalization. However, there exists considerable artifacts in its generations. The α-GAN of Rosca et al. (2019), which integrates the base DCGAN in its framework performs significantly better (31.1 vs 19.2 FID). This shows the effectiveness of hybrid VAE-GAN frameworks in increasing quality and diversity of generations. Our α-GAN + SN regularized with Spectral Normalization performs significantly better (15.9 vs 19.2 FID). Finally, our BMS-VAE-GAN improves significantly over the α-GAN + SN baseline using the "Best-of-Many" samples (14.7 vs 15.9 FID) and slows a clear increase in sharpness.

4 Conclusion

We propose a new objective for training hybrid VAE-GAN frameworks which overcomes key limitations of current hybrid VAE-GANs. We integrate a "Best-of-Many" reconstruction likelihood which helps in covering all the modes of the data distribution while maintaining a latent space as close to Gaussian as possible. Our hybrid VAE-GAN framework outperforms state-of-the-art hybrid VAE-GANs and plain GANs in generative modelling on highly multi-modal synthetic data, CIFAR-10 and CelebA, demonstrating the effectiveness of our approach.

References

J. Adler and S. Lunz. Banach wasserstein gan. *NeurIPS*, 2018.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *ICML*, 2017.

D. Berthelot, T. Schumm, and L. Metz. Began: boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

A. Bhattacharyya, B. Schiele, and M. Fritz. Accurate and diverse sampling of sequences based on a "best of many" sample objective. *CVPR*, 2018.

A. Bozkurt, B. Esmaeili, D. H. Brooks, J. G. Dy, and J.-W. van de Meent. Can vaes generate novel examples? *NeurIPS Workshop*, 2018.

M. Comenetz. *Calculus: the elements*. World Scientific Publishing Company, 2002.

J. Donahue, P. Krähenbühl, and T. Darrell. Adversarial feature learning. *ICLR*, 2017.

V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, and A. Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.

H. Eghbal-zadeh, W. Zellinger, and G. Widmer. Mixture density generative adversarial networks. *CVPR*, 2019.

M. Elfeki, C. Couprie, M. Riviere, and M. Elhoseiny. Gdpp: Learning diverse generations using determinantal point process. *ICML*, 2019.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *NIPS*, 2014.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. *NIPS*, 2017.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *ICLR*, 2014.

A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. *ICML*, 2016.

A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. Adversarial autoencoders. *ICLR*, 2016.

T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. *ICLR*, 2018.

F. Nielsen and K. Sun. Guaranteed bounds on the kullback–leibler divergence of univariate mixtures. *IEEE Signal Processing Letters*, 2016.

A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *ICLR*, 2016.

M. Rosca, B. Lakshminarayanan, , and S. Mohamed. Distribution matching in variational inference. *arXiv preprint arXiv:1802.06847*, 2019.

R. Shu, H. H. Bui, S. Zhao, M. J. Kochenderfer, and S. Ermon. Amortized inference regularization. *NeurIPS*, 2018.

A. Srivastava, L. Valkoz, C. Russell, M. U. Gutmann, and C. Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. *NIPS*, 2017.

J. Tabor, S. Knop, P. Spurek, I. Podolak, M. Mazur, and S. Jastrzebski. Cramer-wold autoencoder. *arXiv preprint arXiv:1805.09235*, 2018.

N.-T. Tran, T.-A. Bui, and N.-M. Cheung. Dist-gan: An improved gan using distance constraints. *ECCV*, 2018.

S. Zhao, J. Song, and S. Ermon. Infovae: Information maximizing variational autoencoders. *arXiv preprint arXiv:1706.02262*, 2017.

Appendix A. Complete derivation

Here we provide complete derivation of (3). For completeness, we begin with a derivation of the multi-sample objective used in Bhattacharyya et al. (2018).

The VAE and hybrid VAE-GANs (Dumoulin et al., 2016; Makhzani et al., 2016; Rosca et al., 2019; Zhao et al., 2017) maximize the log-likelihood of the data ($\mathrm{x} \sim p(\mathrm{x})$). The log-likelihood, assuming the latent space to be distributed according to $p(\mathrm{z})$,

$$\log(p_\theta(\mathrm{x})) = \log\Big(\int p_\theta(\mathrm{x}|\mathrm{z})p(\mathrm{z})dz\Big). \tag{5}$$

Here, $p(\mathrm{z})$ is usually Gaussian. However, the integral in (5) is intractable. VAEs and Hybrid VAE-GANs use amortized variational inference using an (approximate) variational distribution $q_\phi(\mathrm{z}|\mathrm{x})$ (jointly learned using an encoder),

$$\log(p_\theta(\mathrm{x})) = \log\Big(\int p_\theta(\mathrm{x}|\mathrm{z})\frac{p(z)}{q_\phi(\mathrm{z}|\mathrm{x})}q_\phi(\mathrm{z}|\mathrm{x})dz\Big).$$

To arrive at a tractable objective, the standard VAE objective applies the Jensen inequality at this stage, but this forces the final objective to consider the average data-likelihood. Following Bhattacharyya et al. (2018), we apply the Mean Value theorem of Integration (Comenetz, 2002) to leverage multiple samples,

$$\log(p_\theta(\mathrm{x})) \geq \log\Big(\int_a^b p_\theta(\mathrm{x}|\mathrm{z})\, q_\phi(\mathrm{z}|\mathrm{x})\, dz\Big) + \log\Big(\frac{p(\mathrm{z}')}{q_\phi(\mathrm{z}'|\mathrm{x})}\Big),\ \mathrm{z}' \in [a,b]. \tag{6}$$

We can lower bound (6) with the minimum value of z',

$$\log(p_\theta(\mathrm{x})) \geq \log\Big(\int_a^b p_\theta(\mathrm{x}|\mathrm{z})\, q_\phi(\mathrm{z}|\mathrm{x})\, dz\Big) + \min_{\mathrm{z}'\in[a,b]} \log\Big(\frac{p(\mathrm{z}')}{q_\phi(\mathrm{z}'|\mathrm{x})}\Big). \tag{S2}$$

As the term on the right is difficult to estimate, we approximate it using the KL divergence (as in Bhattacharyya et al. (2018)). Intuitively, as the KL divergence heavily penalizes $q_\phi(\mathrm{z}|\mathrm{x})$ if it is high for low values $p(\mathrm{z})$, this ensures that the ratio ${}^{p(\mathrm{z}')}/_{q_\phi(\mathrm{z}'|\mathrm{x})}$ is maximized. Similar to Bhattacharyya et al. (2018), this leads to the "many-sample" objective (4) of the main paper,

$$\mathcal{L}_{\text{MS}} = \log\Big(\mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})}p_\theta(\mathrm{x}|\mathrm{z})\Big) - \mathrm{D}_{KL}(q_\phi(\mathrm{z}|\mathrm{x}) \parallel p(\mathrm{z})). \tag{4}$$

Next, we integrate a synthetic likelihood term with (4). We first convert the likelihood term in (4) to a likelihood ratio form which allows for synthetic estimates, (α, β are regularization terms) by dividing and multiplying $p_\theta(\mathrm{x}|\mathrm{z})$ by $p(\mathrm{x})$,

$$\alpha \log\Big(\mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})}p_\theta(\mathrm{x}|\mathrm{z})\Big) + \beta \log\Big(\mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})}p_\theta(\mathrm{x}|\mathrm{z})\Big) - \mathrm{D}_{KL}(p(\mathrm{z}) \parallel q_\phi(\mathrm{z}|\mathrm{x})) \tag{7}$$

$$\propto\ \alpha \log\Big(\mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})}\frac{p_\theta(\mathrm{x}|\mathrm{z})}{p(\mathrm{x})}\Big) + \beta \log\Big(\mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})}p_\theta(\mathrm{x}|\mathrm{z})\Big) - \mathrm{D}_{KL}(q_\phi(\mathrm{z}|\mathrm{x}) \parallel p(\mathrm{z})).$$

To enable the estimation of the likelihood ratio ${}^{p_\theta(\mathrm{x}|\mathrm{z})}/_{p(\mathrm{x})}$ using a neural network, we introduce the auxiliary variable y where, $\mathrm{y} = 1$ denotes that the sample was generated and $\mathrm{y} = 0$ denotes that the sample is from the true distribution. We can now express (7) as (using Bayes theorem),

$$\begin{aligned} &\alpha \log\Big(\mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})}\frac{p_\theta(\mathrm{x}|\mathrm{z},\mathrm{y}=1)}{p(\mathrm{x}|\mathrm{y}=0)}\Big) + \beta \log\Big(\mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})}p_\theta(\mathrm{x}|\mathrm{z})\Big) - \mathrm{D}_{KL}(q_\phi(\mathrm{z}|\mathrm{x}) \parallel p(\mathrm{z})). \\ =&\alpha \log\Big(\mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})}\frac{p_\theta(\mathrm{y}=1|\mathrm{z},\mathrm{x})}{1-p(\mathrm{y}=1|\mathrm{x})}\Big) + \beta \log\Big(\mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})}p_\theta(\mathrm{x}|\mathrm{z})\Big) - \mathrm{D}_{KL}(q_\phi(\mathrm{z}|\mathrm{x}) \parallel p(\mathrm{z})). \end{aligned} \tag{8}$$

This is because, (assuming independence $p(\mathrm{z},\mathrm{x}) = p(\mathrm{z})p(\mathrm{x})$)

$$p_\theta(\mathrm{x}|\mathrm{z},\mathrm{y}=1) = \frac{p(\mathrm{y}=1|\mathrm{z},\mathrm{x})p(\mathrm{x})}{p(\mathrm{y}=1)}$$

and,

$$p_\theta(\mathrm{x}|\mathrm{y}=0) = \frac{p(\mathrm{y}=0|\mathrm{x})p(\mathrm{x})}{p(\mathrm{y}=0)}.$$

Assuming, $p(\mathrm{y}=0) = p(\mathrm{y}=1)$ (equally likely to be true or generated),

$$\frac{p_\theta(\mathrm{x}|\mathrm{z},\mathrm{y}=1)}{p(\mathrm{x}|\mathrm{y}=0)} = \frac{p_\theta(\mathrm{y}=1|\mathrm{z},\mathrm{x})}{p(\mathrm{y}=0|\mathrm{x})}.$$

Thus, finally we have,

$$= \alpha \log\left(\mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})} \frac{p_\theta(\mathrm{y}=1|\mathrm{z},\mathrm{x})}{p(\mathrm{y}=0|\mathrm{x})}\right) + \beta \log\left(\mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})} p_\theta(\mathrm{x}|\mathrm{z})\right) - \mathrm{D}_{KL}(q_\phi(\mathrm{z}|\mathrm{x}) \parallel p(\mathrm{z}))$$

$$= \alpha \log\left(\mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})} \frac{p_\theta(\mathrm{y}=1|\mathrm{z},\mathrm{x})}{1-p(\mathrm{y}=1|\mathrm{x})}\right) + \beta \log\left(\mathbb{E}_{q_\phi(\mathrm{z}|\mathrm{x})} p_\theta(\mathrm{x}|\mathrm{z})\right) - \mathrm{D}_{KL}(q_\phi(\mathrm{z}|\mathrm{x}) \parallel p(\mathrm{z})).$$

This completes the derivation.

Appendix B. Training Algorithm

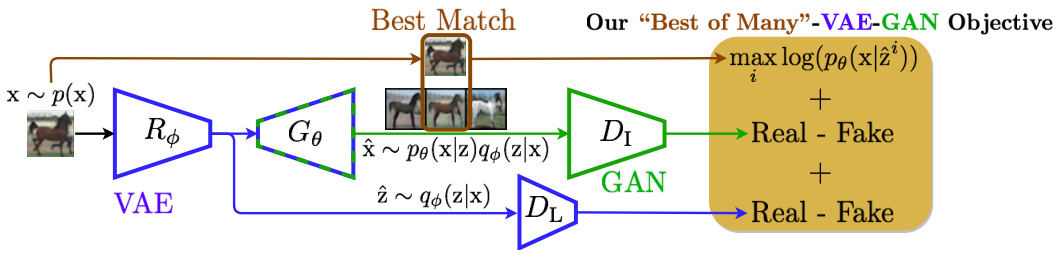


Figure 2: Overview of our BMS-VAE-GAN architecture at training time. The terms of our novel objective (4) are highlighted at the right. We consider only the best sample from the generator G_θ while computing the reconstruction loss. This enables us to generate diverse samples covering all modes of the data distribution while maintaining the low divergence to the prior.

Algorithm 1: BMS-VAE-GAN Training.

Initialize parameters of $R_\phi, G_\theta, D_\mathrm{I}, D_\mathrm{L}$;
for $i \leftarrow 0$ **to** *max_iters* **do**
  Update R_ϕ, G_θ (jointly) using our $\mathcal{L}_{\text{BMS-S}}$ objective;
  Update D_I using hinge loss to produce high values ($\geq a$) for real images and low ($\leq b$) otherwise: $\mathbb{E}_{p(\mathrm{x})} \max\{0, a - \log(D_\mathrm{I}(\mathrm{x}))\} + \mathbb{E}_{p(\mathrm{z})} \max\{0, b + \log(D_\mathrm{I}(G_\theta(\mathrm{z})))\}$;
  Update D_L using the standard cross-entropy loss: $\mathbb{E}_{p(\mathrm{z})} \log(D_\mathrm{L}(\mathrm{z})) + \mathbb{E}_{p(\mathrm{x})} \log(1 - D_\mathrm{L}(R_\phi(\mathrm{x})))$;
end

We detail in algorithm 1, how the components $R_\phi, G_\theta, D_\mathrm{I}, D_\mathrm{L}$ of our BMS-VAE-GAN (see Figure Figure 2) are trained. We follow Rosca et al. (2019) in designing algorithm 1. However, unlike Rosca et al. (2019), we train R_ϕ, G_θ jointly as we found it to be computationally cheaper without any loss of performance. Unlike Rosca et al. (2019), we use the hinge loss to update D_I as it leads to improved stability (as discussed in the main paper).

Appendix C. Qualitative Examples on CelebA

We provide qualitative results on CelebA in Figure 3. We observe in Figure 3a that although the SN-GAN (Miyato et al., 2018) produces sharp images, we also observe considerable number of artifacts. In contrast, the generations of our α-GAN + SN do not contain such artifacts. Finally, the addition of the "Best of Many" sample loss leads to a considerable increase in sharpness.

(a) SN-GAN (Miyato et al., 2018)

(b) α-GAN + SN ($T = 1$) (Rosca et al., 2019)

(c) Our BMS-VAE-GAN ($T = 30$)

Figure 3: Qualitative results on CelebA.