NATURAL SCENE STATISTICS IN ADVERSARIAL LEARNING

Pavan Chennagiri (UT EID: pm29224)

Electrical and Computer Engineering,
The University of Texas at Austin, Austin, Texas, USA
pavancm@utexas.edu

ABSTRACT

Synthesizing natural images using generative models such as Generative Adversarial Network (GAN) has received significant attention in the recent days due to advancements in deep learning. The existing generative models employ relatively simple loss functions derived from L_1/L_2 norms during training due to it's simplistic nature as well very desirable properties in statistics and estimation. However from the perceptual viewpoint mean squared error (L_2 norm) has a very weak correlation with image quality. In this work the effect of incorporating statistics that effectively quantify the 'naturalness' of an image is studied. In particular distances derived from Natural Scene Statistics is used as a constraint while learning the generative model. Specifically the performances of Multi-scale Structural Similarity (MS-SSIM) and Visual Information Fidelity (VIF) and their advantages as well as shortcomings are holistically analyzed.

Index Terms— Generative Adversarial Networks, Boundary-equilibrium, autoencoders, image quality assessment, structural similarity index, visual information fidelity

1. INTRODUCTION

Generative Adversarial Networks are a class of unsupervised generative models which learn the data distribution p_{data} using a large corpus of data by means of an adversarial loss. GANs were first proposed in [1] and they are built around two functions: the generator $G(z)$ generates a data sample from p_{data} for a random z sampled from a uniform distribution, (z is also referred to as latent input) and the discriminator $D(x)$ provides inference whether the sample x belongs to the data distribution p_{data}. From game theoretic standpoint GANs are viewed as a minimax zero sum game played between two players (the two players here are G and D) and the learning occurs in a joint fashion where D and G train in an alternate manner.

Since their introduction various GAN architectures have been proposed in the literature. Deep Convolutional GAN (DCGAN) [2] used Convolutional Neural Networks (CNN) both in G and D. DCGAN along with the original GAN used Jenson-Shannon (JS) Divergence which is a symmetric version of Kullback Leibler (KL) Divergence in their objectives while training for measuring the distance between the model distribution p_{model} and p_{data}. However JS divergence is known to suffer from vanishing gradient problem where gradients decay to zero resulting in no learning. Wasserstein GAN (WGAN) [3] addressed this problem by employing Wasserstein distance in place of JS Divergence. Although WGAN provided better performance it came at the expense of slow training. In Energy Based GANs (EBGANs) [4] the discriminator D is modeled as an energy function by means of an auto-encoder and has been shown to be easier to train as well as generate better looking images. Boundary Equilibrium GAN (BEGAN) [5] is motivated from EBGAN but instead of matching the data distributions it aims to match loss distribution of the autoencoder. In this work I restrict all the experimental analysis to the BEGAN model.

The main challenges in training GANs stem from the fact there exists no straightforward way in determining the hyperparameters. The convergence of GANs is strictly tied to particular choice of hyperparameters which are very sensitive to the choice of dataset as well as GAN architecture. Also GANs easily suffer from mode collapse where it learns to generate a single image for every latent input z. Also the GAN objectives are typically multimodal where convergence does not necessarily translate to generating a naturalistic image. The existing GAN objectives do not employ any measure of naturalness in their objectives and typically use loss functions which have been shown to have good statistical properties for convergence. To enforce 'naturalness' in generated images [6] proposed employing Multi-Scale Structural Similarity Index (MS-SSIM) in GAN objective. In this work an effort is made to explicitly incorporate Natural Scene Statistics (NSS) in GANs and analyze it's behavior with regard to improving the naturalness of generated images.

2. MODEL DESCRIPTION

This section provides a brief overview of BEGAN model. In the original GAN method [1] GAN objective was a minimax

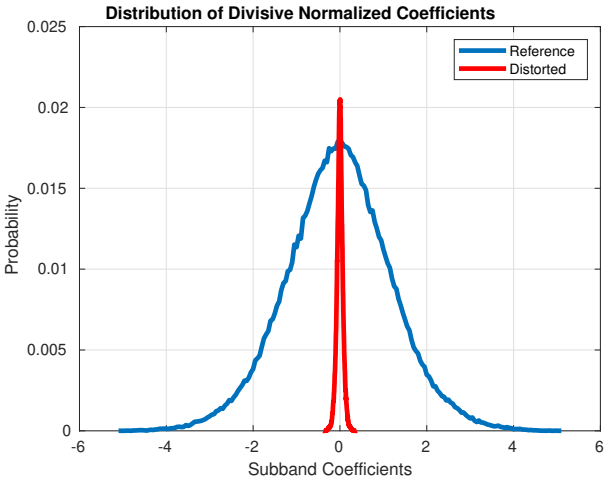


Fig. 1: Comparison of distribution of divisive normalized coefficients. In the above plot the distribution of Gaussian blur distortion is compared with that of the reference

function shown in equation 1

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data(x)}}[\log(D(x))] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))], \quad (1)$$

where p_{data} is the true data distribution and $p_z(z)$ is the distribution from which sample z is drawn from. In case of BEGAN the discriminator is replaced by an auto-encoder. If $x \in R^{N_x}$, then $D(x) : R^{N_x} \rightarrow R^{N_x}$, BEGAN aims to match auto-encoder loss distributions using Wasserstein (earth-mover) distance. The loss function for the auto-encoder is given by

$$L(x) = |x - D(x)|^n; n = 1, 2, \quad (2)$$

where $L(x)$ is the loss of real image sample x and $L(G(z))$ is the corresponding loss of generated sample $G(z)$. Since the aim is to match the loss functions with respect to Wasserstein distance, upon modification GAN objective reduces to

$$\begin{aligned} L_D &= L(x) - k_t L(G(z)) \\ L_G &= L(G(z)) \\ k_{t+1} &= k_t + \lambda_k(\gamma L(x) - L(G(z)), \end{aligned} \quad (3)$$

where $\gamma \in [0, 1]$ is a hyperparameter referred to as diversity ratio mathematically represented as

$$\gamma = \frac{\mathbb{E}[L(G(z))]}{L(x)}, \quad (4)$$

where γ term balances between two goals: auto-encode real images and discriminate real from generated images. Lower values of γ gives more prominence to auto-encoding while larger values result in more diversified image model. The

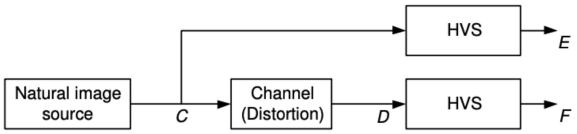


Fig. 2: Human Visual System model employed in VIF

term $k_t \in [0, 1]$ is a parameter at each iteration step in order to maintain the equilibrium $\mathbb{E}[L(G(z))] = \gamma\mathbb{E}[L(x)]$. The network architecture employed in BEGAN is shown in Fig. 4.

3. NATURAL SCENE STATISTICS

The loss function defined in equation 2 translates to either Mean Squared Error (MSE) or Mean Absolute Difference (MAD). From the image quality assessment literature it has been well studied that both of these metrics are weakly correlated with human perception. In other words the auto-encoder loss does not explicitly require the solution to be an image that follows statistics associated with natural images. A simple fix will be to add an additional constraint in the form of loss function. Mathematically it can be written as

$$L(x) = \lambda_1 L_1(x) + \lambda_2 L_2(x), \quad (5)$$

where $L_1(x)$ is MSE/MAD and $L_2(x)$ is a measure of deviation of generated images from that of natural images. In [6] MS-SSIM was used as $L_2(x)$. In this project a more fundamental way of measuring deviation is studied.

It has been well studied in the perception literature that natural images follow certain behavior which is not observed in distorted/synthesized images [7]. For instance, it is widely observed that the average power spectrum magnitude of a natural image follows an approximate inverse square law with respect to frequency. Also another widely used observation is that the distribution of band pass coefficients of natural images follow a heavy tailed distribution implying the non-Gaussian nature of natural images in band-pass/wavelet domain. The latter has been successfully used in image quality metrics such as BRISQUE [8], NIQE [9] etc. A simple way to find deviation between the distributions is to measure the KL Divergence. In [10] a Gaussain Scale Mixture (GSM) based transform called divisive normalization (DN) was proposed which tries to gaussianize the distribution of subband coefficients. However the same argument does not hold in case of distorted images, thus the deviation from the reference distribution is an indicator of quality. If the subband is denoted by Y, then by GSM model $Y \equiv sU$, where $\equiv$ denotes equality in distribution, $s \geq 0$ is a scalar random variable and U is a Gaussian random vector with zero mean and known covariance C_U. It is also assumed that s and U are independent. In case of subbands, subband coefficients constitute Y, C_U is

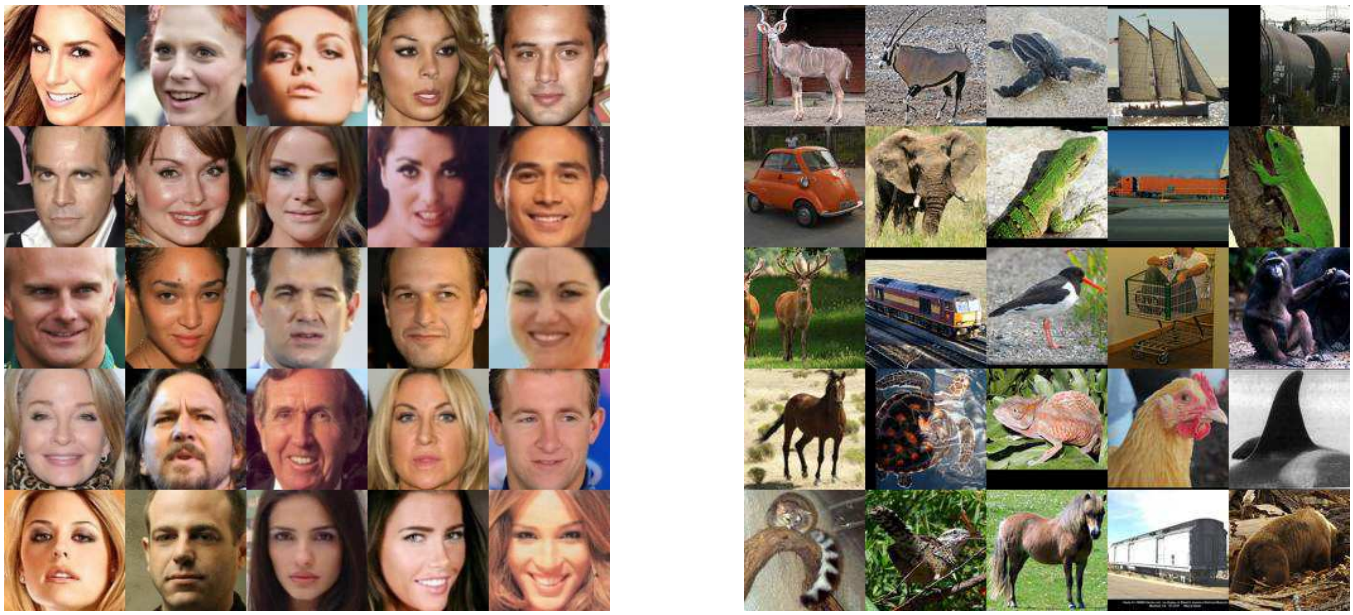


Fig. 3: Sample images from CelebA (left) and STL-10 (right) datasets

empirically computed and s is calculated through maximum likelihood estimation (MLE) given by

$$\hat{s} = \arg\max_{z} \log(p(Y/s)) = \sqrt{\frac{Y^T C_U^{-1} Y}{N}} \tag{6}$$

An illustration of deviation observed in distribution of divisive normalized coefficients is shown in Fig. 1 where the distribution of Gaussian blur distorted image is compared with that of reference image.

Calculating KL divergences poses practical challenges. Often calculating the density functions for every image is cumbersome and inaccurate which affects the calculated value. And KL divergence is not symmetric as well as unbounded, the latter has a profound effect in the initial stages of GANs where the distribution of generated images significantly deviate from that of natural images resulting in very large KL divergence, thus affecting the learning process. Taking motivations from information theory, Visual Information Fidelity (VIF) [11] employs mutual information in place of KL divergence to measure the quality and has been shown to perform well in a variety of tasks. In this work the effect of using VIF in a GAN scenario is studied.

A brief description of VIF is provided here. VIF is motivated from information theoretic standpoint of human visual system as shown in Fig. 2 where the distorted image is equivalent to an image passing through a noisy communication channel. All the signals shown in Fig. 2 are assumed to be in wavelet domain with C following a GSM model

$$C \equiv sU \tag{7}$$

$$D \equiv gC + V \tag{8}$$

where g is a deterministic gain factor and V is a zero mean stationary additive Gaussian noise with covariance $C_V = \sigma_v^2 \mathbf{I}$. Here g and C_V characterize the nature of distortion. Also

$$E = C + N \tag{9}$$

$$F = D + N', \tag{10}$$

where N and N' constitute noise introduced by the visual system. Here as well for simplicity it's assumed that N and N' are zero mean Gaussians with covariance $C_N = C_{N'} = \sigma_n^2 \mathbf{I}$. The difference between the mutal information $I(C;E)$ and $I(C;F)$ is an indicator of the distortion suffered which in turn leads to measuring image quality. Using properties of mutual information the expression can be simplified as,

$$I(C;E) = \frac{1}{2} \sum_{i=1}^{N} \log \left(\frac{|s_i^2 C_U + \sigma_n^2 \mathbf{I}|}{|\sigma_n^2 \mathbf{I}|} \right) \tag{11}$$

$$I(C;F) = \frac{1}{2} \sum_{i=1}^{N} \log \left(\frac{|g_i^2 s_i^2 C_U + (\sigma_n^2 + \sigma_v^2) \mathbf{I}|}{|(\sigma_n^2 + \sigma_v^2) \mathbf{I}|} \right), \tag{12}$$

where $|.|$ denotes determinant, C_U, σ_n, σ_v are empirically calculated using given reference and distorted images while g is calculated using linear regression of equation 8 for every wavelet coefficient. Finally the VIF score is obtained as a

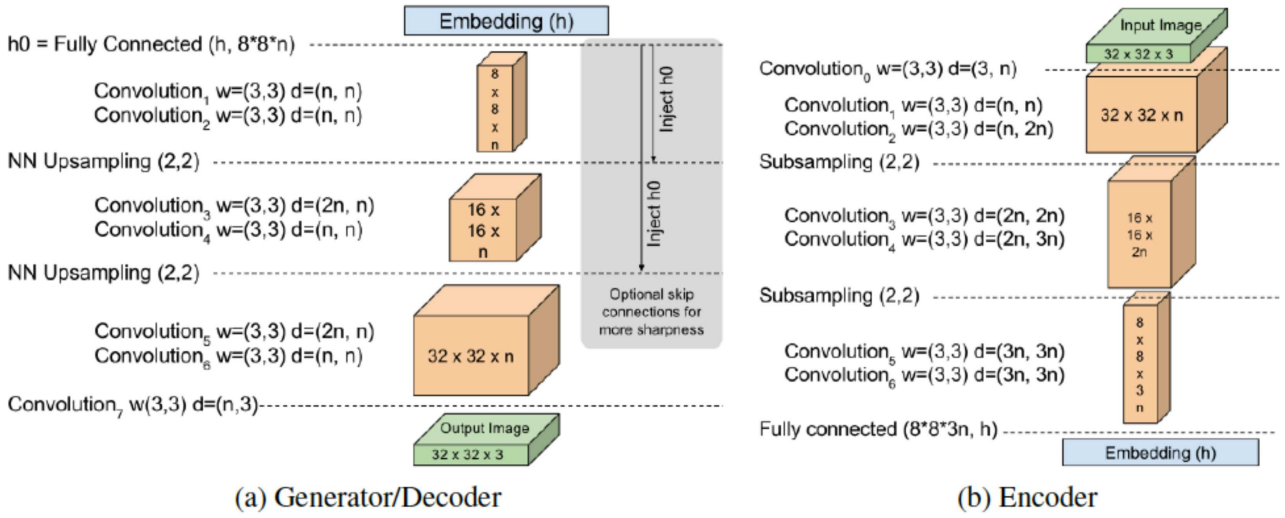


Fig. 4: Network architecture of generator and discriminator in BEGAN. Image obtained from [5]

ratio calculated across different subbands

$$VIF = \frac{\sum_{j \in \text{subbands}} I(C^j; F^j)}{\sum_{j \in \text{subbands}} I(C^j; E^j)} \tag{13}$$

Since distortion introduces information loss, the mutual information $I(C; E) \geq I(C; F)$. Thus $VIF \in [0, 1]$.

4. MODEL DESCRIPTION

This section contains the details of the model as well as the metrics employed to holistically evaluate the trained model. Also a brief description of the datasets used for training and the parameters used while training is provided.

For constraining the images to be more 'natural', the VIF is employed in $L_2(x)$ term in equation 5 as

$$L_2(x) = 1 - \text{VIF}(x, D(x)), \tag{14}$$

as high quality images have larger VIF values and thus result in lower loss values. VIF is calculated on subbands obtained by decomposing the given image into steerable pyramids [12]. Also $L_2(x) = 1 - \text{MS-SSIM}(x, D(x))$ proposed in [6] is used to compare the performance with VIF loss. All the experiments are conducted by keeping the architecture of GAN constant, which in this case is the BEGAN architecture as shown in Fig. 4.

4.1. Dataset

Training GANs necessitate the need of large quantum of data. In this work the GANs are evaluated on CelebA [13] and STL-10 [14] datasets. CelebA consists of more than 200,000 face images of celebrities. The data was pre-processed by employing a face detection algorithm and images which had low confidence in detecting faces were rejected. The selected images were then downsampled to 128×128 and linearly scaled to lie in $[-0.5, 0.5]$.

STL-10 dataset consists of more than 100,000 images captured across diverse scenes, all of which have a resolution 96×96. In case of STL-10 as well, the images were linearly rescaled to lie in $[-0.5, 0.5]$. Sample images from CelebA and STL-10 databases are shown in Fig. 3.

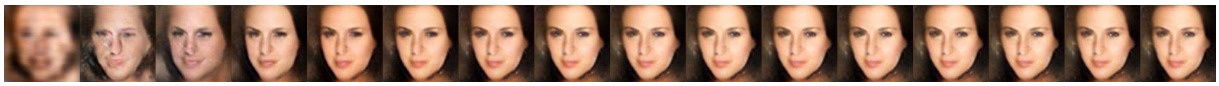
I used the default parameters recommended by the authors in [5] such as $\lambda_k = 0.001, k_0 = 0, \gamma = 0.7$ which are used in the update expression in equation 3. The models were trained for approximately 100,000 iterations with a batch size of 16. The steerable subbands used in computing VIF in equation 14 were selected as recommended by the authors in [11].

4.2. Training

Training was conducted separately for CelebA and STL databases. The dimension of the latent input z in both the cases was fixed to 64 while the number of filters in the decoder/encoder (in Fig. 4 the variable n corresponds to number of filters) was fixed to 128 as recommended by the authors in [5]. The number of filters define the complexity of the model. The choice of number of filters is a trade-off where having large number of filters can make it hard to train due to higher number of parameters and possibly overfit the data, while having lower number of filters can make the model simplistic in nature with the model not learning the intricacies of the training data. All the models were trained using a Nvidia Tesla P100 GPU, with the original BEGAN and BEGAN with MS-SSIM loss requiring approximately 18 hours to train for 100,000 iterations, while BEGAN with VIF loss required more than 40 hours to do the same. The significantly longer time for model with VIF loss is attributed to the complex operations involved in computing VIF (particularly involving

(a) BEGAN

(b) BEGAN with MS-SSIM loss



(c) BEGAN with VIF loss

Fig. 5: Image generation progression of BEGAN with different loss functions across iterations for CelebA database. The iterations increase from left to right. More images are attached in Appendix A

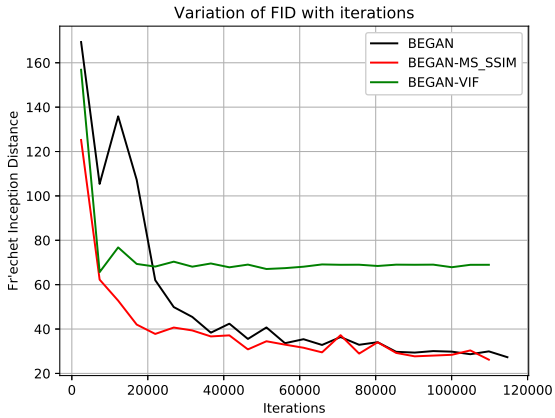


Fig. 6: Variation of FID with iterations for CelebA dataset

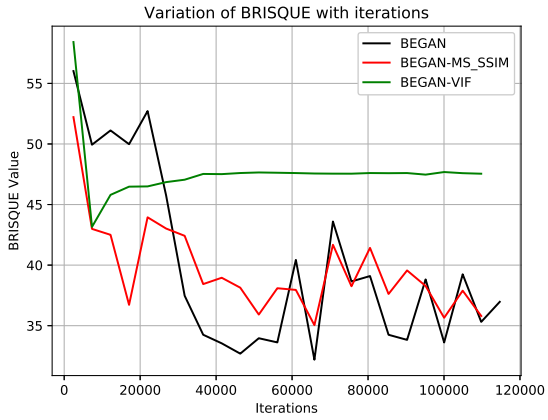


Fig. 7: Variation of BRISQUE with iterations for CelebA dataset

steerable pyramid decomposition as well as matrix inversion given in equation 6). In all the above cases mini-batch gradient descent with Adam optimization [15] was employed with a batch size of 16 images per iteration. All the experiments were conducted with loss functions $L_1(x)$ and $L_2(x)$ in equation 5 receiving equal weightage with $\lambda_1 = \lambda_2 = 0.5$ and $L_1(x)$ being equal to L_1 norm.

4.3. Evaluation Methodology

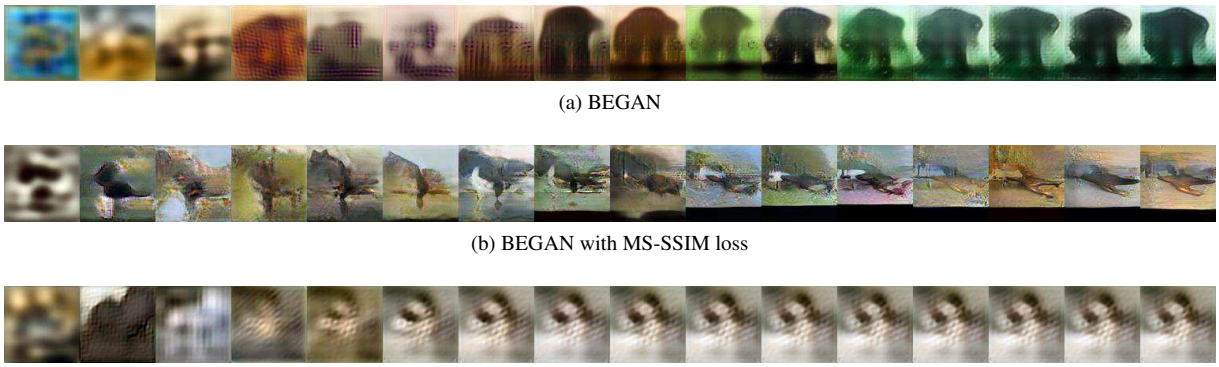
For comparing the performance of the proposed approach with other loss functions, a metric called Fréchet Inception Distance (FID) [16] is used. FID is a statistically inspired technique which measures the quality of generated samples using Inception network [17] to compute the features from a specific layer. The generated images are fed to an inception model that was trained on Imagenet database and features are obtained from a particular layer. These features are empirically observed to follow a multivariate Gaussian distribution for large class of images. The mean and covariance are calculated for both real data as well as generated data using these computed features and the distance between two Gaussians is measured by Fréchet distance [18] (also known as Wasserstein-2 distance) in order to quantify the quality of the generated samples.

	BEGAN	BEGAN-MS-SSIM	BEGAN-VIF
FID	27.30	26.18	65.65
BRISQUE	56.01	52.21	58.41

Table 1: Quality scores for different BEGAN methods for CelebA dataset.

$$FID(x, g) = ||\mu_x - \mu_g||_2^2 + \text{Tr}(\Sigma_x + \Sigma_g - 2(\Sigma_x\Sigma_g)^{\frac{1}{2}}), \tag{15}$$

where (μ_x, Σ_x) and (μ_g, Σ_g) are the mean and covariance matrix of features obtained from real and generated data respectively. FID scores are negatively correlated with visual quality where lower distance imply that the distribution of



(a) BEGAN

(b) BEGAN with MS-SSIM loss

(c) BEGAN with VIF loss

Fig. 8: Image generation progression of BEGAN with different loss functions across iterations for STL-10 database. The iterations increase from left to right. More images are attached in Appendix B

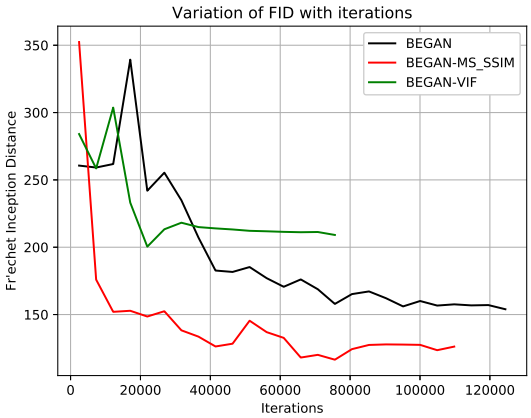


Fig. 9: Variation of FID with iterations for STL-10 dataset.

Fig. 10: Variation of BRISQUE with iterations for STL-10 dataset.

	BEGAN	BEGAN-MS-SSIM	BEGAN-VIF
FID	153.93	116.47	200.42
BRISQUE	52.41	47.74	54.41

Table 2: Quality scores for different BEGAN methods for STL-10 database.

generated images is closer to that of the real data, thus higher quality.

I also used Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [8] score for evaluating the quality of generated images. BRISQUE is a popular no-reference image quality assessment metric derived from NSS and measure the quality based on the deviation observed in the distribution of contrast normalized coefficients when compared to that of natural images. BRISQUE provides values lying in $[0, 100]$ with better quality images receiving higher values. Thus BRISQUE is positively correlated with image quality which is in contrast to that of FID.

5. EXPERIMENTS AND RESULTS

This section contains the comparisons in the performance of three models BEGAN, BEGAN with MS-SSIM and BEGAN with VIF loss functions across three different quality metrics detailed in the previous section. The comparison is made across two datasets CelebA and STL-10.

The progression of BEGAN training for CelebA across the number of iterations is illustrated in Fig. 5. It can be clearly observed that in the initial stages the generated images are highly distorted, but as the iterations increase the images become progressively better. The same is illustrated objectively in Figs. 6 and 7 where FID and BRISQUE scores are plotted across iterations respectively. From FID plot it can be concluded that BEGAN-VIF has inferior performance when compared to other methods although in the initial stages of training the improvement in performance is similar to MS-SSIM, it saturates after 20,000 iterations with negligible change in quality of generated images. This observation

translates to the loss of details in the generated images obtained from BEGAN-VIF (This can be observed from Fig. 5 as well as in Appendix A). A possible explanation for this behavior can be that employing NSS alone might be insufficient in capturing all the attributes of natural images. Although the loss function $L(x)$ consists of both $L_1(x)$ corresponding to L_1 norm and $L_2(x)$ derived from NSS, the performance drastically reduces for $L_2(x) = 1 - VIF(x, D(x))$ when compared to that of MS-SSIM. One possible way to explain this is that VIF does not explicitly measure the details (the presence/absence of edges) of the image but it only characterizes the deviation in the distribution as a whole. So the facial expressions which are perceptually significant appear to have no influence on the predicting VIF score thus resulting in a saturating behavior.

Observing variation of BRISQUE in Fig. 7 leads to contradictory implications as BEGAN-VIF seemingly has a higher value than other models although visually images generated from other models have better quality. Also the variation of BRISQUE values is restricted to only between $[35, 55]$. These two observations indicate that BRISQUE which is itself a trained model might not be capturing the distortions that is observed in these generated images. The BRISQUE is trained on images suffering from commonly observed distortions such as blur, JPEG etc. However in GAN images the artifacts are more geometric in nature having unnatural shapes and structures. Table 1 lists the best objective score obtained for different models.

The progress of BEGAN training in case of STL-10 database in Fig. 8. It can be observed from the figure that the quality of generated images is significantly poorer than that of CelebA dataset. This can be attributed to the fact that the STL-10 dataset consists of a wide variety of scenes as opposed to just face images in CelebA. In other words CelebA has more well defined structure in the form of faces as opposed to diverse images in STL-10 requiring the generative model to be more complex for encapsulating wide variation in scenes. Also BEGAN as well as BEGAN-VIF fails to sufficiently converge to the optimal saddle point of the GAN objective leading to mode collapse where the generated images are same regardless of latent inputs as shown in Fig. 8. This indicates that the architecture that has a superior performance for a particular dataset might not achieve similar performance when employed on a different data implying the data sensitiveness of the trained model.

Fig. 6 shows the variation of FID with iterations. it follows the observation that BEGAN and BEGAN-VIF didn't sufficiently converge thus resulting in higher FID values. In case of BRISQUE shown in Fig. 10 the observation is similar to that observed for CelebA with BEGAN-VIF having larger values although visually generated images have poor quality. Table 1 lists the best objective score obtained for different models for STL-10 data.

6. DISCUSSION AND CONCLUSION

Using two player zero sum game strategy for image generation task produces images that seemingly appear 'unnatural' due to perceptual artifacts. The existing GAN frameworks do not incorporate any perceptual quality constraint in their model thus leading to images not necessarily following the statistics of natural images. In this work an attempt is made in constraining the model to follow NSS by means of employing a metric measuring the deviation of the distributions between the natural and distorted images. Towards this end the quality metric VIF was used in the training procedure of BEGAN and it's performance was compared to the original BEGAN as well as BEGAN with MS-SSIM model. The motivation to employ VIF arises from it's treatment of measuring the quality as a difference in the mutual information, which in turn is closely related to KL Divergence. Although VIF imposes NSS constraints, performance wise it fails to achieve better quality images when compared with MS-SSIM model. It specifically performs poorly in generating intricate details which have profound influence on the quality particularly in facial images.

There are some fundamental limitations to all the above models discussed in this work. Firstly the above models are very sensitive to training data. In this work it was seen that a model with superior performance on facial images has a drastic performance reduction in image quality when trained on a different class of images. This significantly restricts the generalisability of the model across different classes of data. Secondly the models are very sensitive to hyperparameters, particularly the learning rate and diversity ratio used in equations 3 and 4. Even minor variations in these values can affect the convergence of these models leading to mode collapse. Additionally the choice of these parameters are often random with no particular justification reasoning the achieved performance.

7. FUTURE WORK

There are two different arenas that could be looked upon. First is to improve generated image quality. This can be done either by arriving at a different architecture or by imposing perceptual quality constrains in a more direct fashion instead of just using in the loss function. One possible approach can be imposing certain constraints at every layer of the GAN. Second is to design a better metric to objectively quantify the quality of generated images. It was seen in this work that BRISQUE performed poorly in deciding the quality and often had no correlation with observed visual quality. Although FID performed well it is inherently biased towards a particular data (in this case it's ImageNet database) since it employs a trained Inception model for providing quality scores.

8. REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[2] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[3] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning*, 2017, pp. 214–223.

[4] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based generative adversarial network," *arXiv preprint arXiv:1609.03126*, 2016.

[5] D. Berthelot, T. Schumm, and L. Metz, "BEGAN: Boundary equilibrium generative adversarial networks," *arXiv preprint arXiv:1703.10717*, 2017.

[6] P. Kancharla and S. S. Channappayya, "Improving the visual quality of generative adversarial network (GAN)-generated images using the multi-scale structural similarity index," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 3908–3912.

[7] W. S. Geisler, "Visual perception and the statistical properties of natural scenes," *Annu. Rev. Psychol.*, vol. 59, pp. 167–192, 2008.

[8] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[9] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.

[10] M. J. Wainwright and E. P. Simoncelli, "Scale mixtures of gaussians and the statistics of natural images," in *Advances in neural information processing systems*, 2000, pp. 855–861.

[11] H. R. Sheikh, A. C. Bovik, and G. De Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on image processing*, vol. 14, no. 12, pp. 2117–2128, 2005.

[12] E. P. Simoncelli and W. T. Freeman, "The steerable pyramid: A flexible architecture for multi-scale derivative computation," in *Proceedings., International Conference on Image Processing*. IEEE, 1995, vol. 3, pp. 444–447.

[13] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.

[14] A. Coates, A. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 2011, pp. 215–223.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

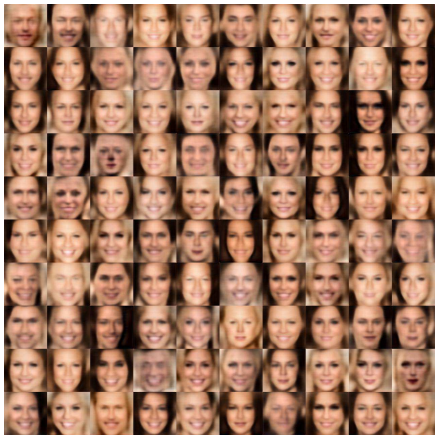
[16] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 6626–6637.

[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

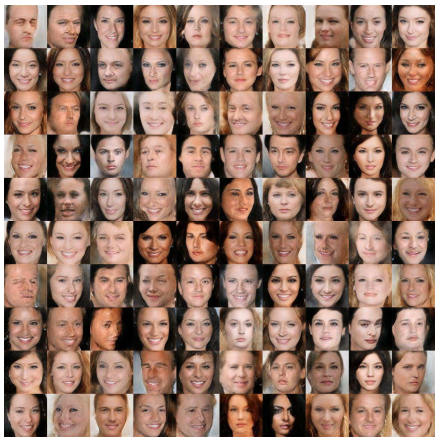
[18] M. Fréchet, "Sur la distance de deux lois de probabilité," in *C. R. Acad. Sci. Paris*, 1957, pp. 244:689–692.

A. CELEBA DATABASE RESULTS

In Figs. 11,12 and 13 the generated images at three different iterations of training is illustrated. It can be clearly observed that in the initial stages the generated images are poor quality while the quality improves upto certain number of steps beyond which it saturates. Also the visual image quality obtained with BEGAN-VIF loss is comparatively poor when



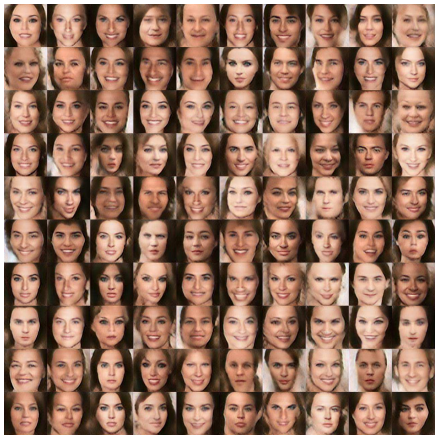
(a) iter -7319



(b) iter -36599

(c) iter -100039

Fig. 11: Images generated at different iterations of training with BEGAN



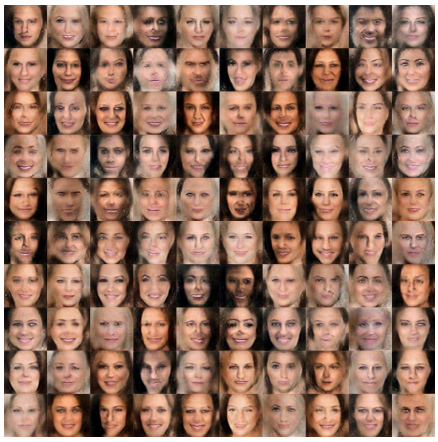
(a) iter -7319

(b) iter -36599

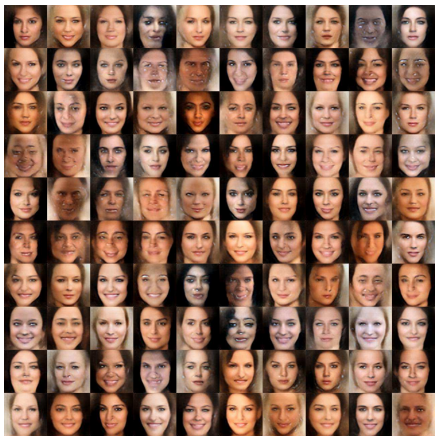
(c) iter -100039

Fig. 12: Images generated at different iterations of training with BEGAN with MS-SSIM loss

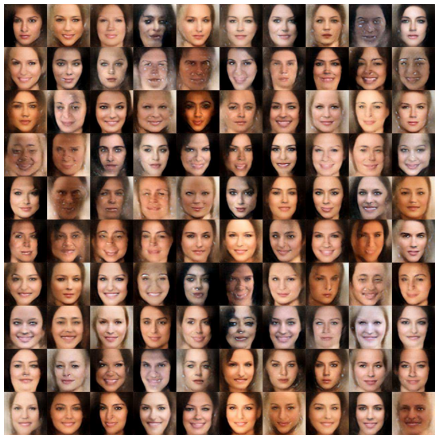
compared to other methods. (The images can be zoomed in to clearly observe the image quality)



(a) iter -7319



(b) iter -36599

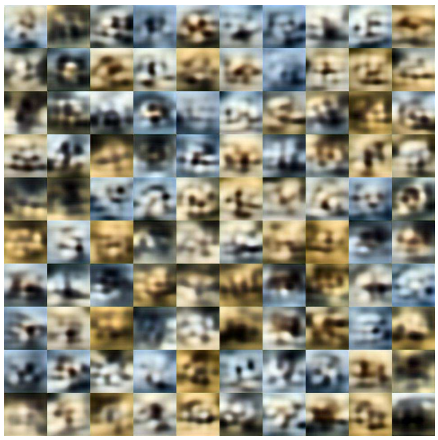


(c) iter -100039

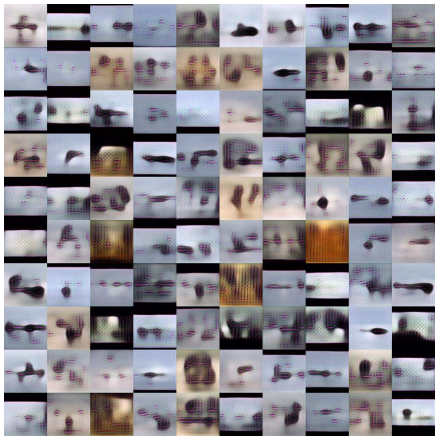
Fig. 13: Images generated at different iterations of training with BEGAN with VIF loss

B. STL-10 DATABASE

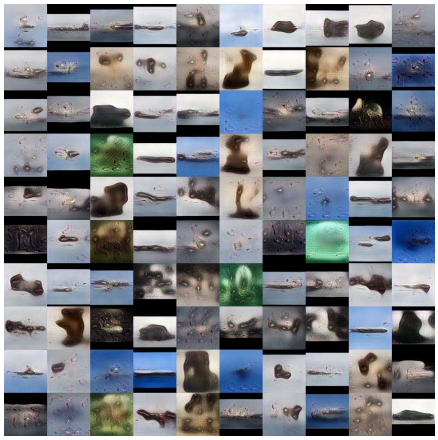
STL-10 database consists a diverse category of scenes. The results in the initial iterations appear to be random noisy im-



(a) iter -12199



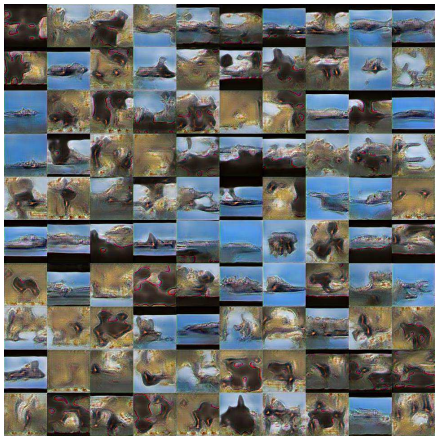
(b) iter -31719



(c) iter -70759

Fig. 14: Images generated at different iterations of training with BEGAN

ages but the quality doesn't improve significantly with more steps as was the case in CelebA dataset. This is illustrated in Figs. 14, 15 and 16. they contain random structures which appear to be quite unnatural. (Images can be zoomed for better clarity)

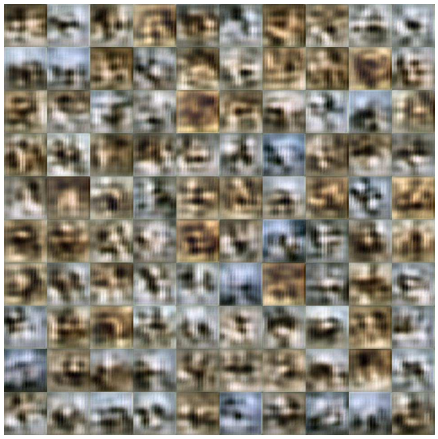


(a) iter -12199

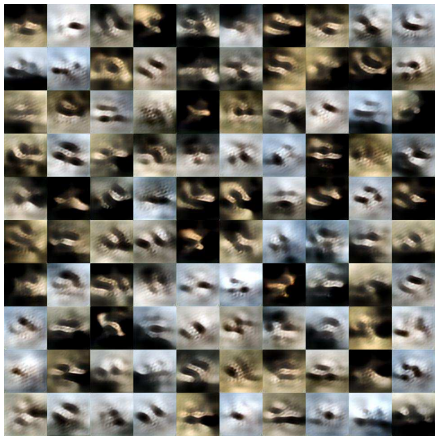
(b) iter -31719

(c) iter -70759

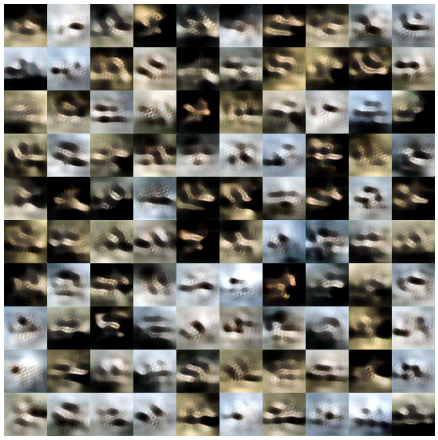
Fig. 15: Images generated at different iterations of training with BEGAN with MS-SSIM loss



(a) iter -12199



(b) iter -31719



(c) iter -70759

Fig. 16: Images generated at different iterations of training with BEGAN with MS-SSIM loss