Quality Map Fusion for Adversarial Learning

Uche Osahor
uo0002@mix.wvu.edu

Nasser M. Nasrabadi
nasser.nasrabadi@mail.wvu.edu

West Virginia University
Morgantown, USA.

Abstract

Generative adversarial models that capture salient low-level features which convey visual information in correlation with the human visual system (HVS) still suffer from perceptible image degradations. The inability to convey such highly informative features can be attributed to mode collapse, convergence failure and vanishing gradients. In this paper, we improve image quality adversarially by introducing a novel quality map fusion technique that harnesses image features similar to the HVS and the perceptual properties of a deep convolutional neural network (DCNN). We extend the widely adopted l_2 Wasserstein distance metric to other preferable quality norms derived from Banach spaces that capture richer image properties like structure, luminance, contrast and the naturalness of images. We also show that incorporating a perceptual attention mechanism (PAM) that extracts global feature embeddings from the network bottleneck with aggregated perceptual maps derived from standard image quality metrics translate to a better image quality. We also demonstrate impressive performance over other methods.

1 Introduction

A representation of the human visual system (HVS) is necessary to establish a robust image quality metric which is needed for computer vision applications [6, 27]. Classical approaches considered hand-crafted strategies to mimic the properties of the HVS [42, 44, 46, 51] by implementing a stream of computational functions that are combined to identify the key perceptual properties of images. While these techniques played their role effectively, scaling up these methods have proven to be a daunting task, especially for applications with huge datasets. However, the introduction of neural networks has helped to improve the aforementioned task considerably [2, 8, 12, 14, 18, 50, 52]. These deep networks consist of non-linear filters configured to extract key perceptual features within user-defined constraints from data.

In our work, we focus more on Full-Reference Image Quality Assessment (FR-IQA) models [23, 52], these models are mostly used to represent the HVS with the aim of deriving a quality measure for images by comparing the perceptual similarity between distorted images and their respective reference image. A standard FR-IQA model seeks to imitate the HVS by exploiting photographic computational algorithms that represent contrast sensitivity, visual masking, luminance, etc. A number of FR-IQA metrics have since being derived which include the Structure Similarity Index Measure (SSIM) [44], Multi Scale Structure Similarity Index Measure (MS-SSIM) [53], Visual Information Fidelity (VIF) [10], Feature Similarity Index Measure (FSIM) [51], Mean Deviation Similarity Index (MDSI) [51], etc.

In this paper, we present a novel approach for improving the quality of GAN-synthesised images by combining the benefits of established FR-IQA metrics and the features of a deep

convolutional neural network (DCNN). We extend the performance of popular GAN-based baseline approaches by introducing a novel image quality map fusion network that computes the perceptual properties of images and fuse them with a perceptual attention mechanism, as shown in Figure 1. We also introduce novel quality loss functions derived via Banach spaces to boost image quality. Our technique shows impressive results as compared to state-of-the-art. Our key contributions are as follows:

- We introduced a new perceptual quality map fusion network that harnesses the perceptual qualities of computationally derived quality assessment metrics.

- We propose a new norm implemented via Banach Wasserstein GAN (BWGAN) instead of the popular l_2 norm computed using the Wasserstein metric.

- We also propose a perceptual attention mechanism (PAM) that augments image features to boost the overall visual appeal of the synthesised images.

2 Related Work

The FR-IQA model tries to simulate the HVS characteristics with good performance measures [35, 36, 48, 53]. Two main reasons for the success of FR-IQA can be attributed to the deep learning based perceptual properties of the reference image and the hand-crafted features derived from statistical metrics which are similar to the HVS. Hence, it becomes easier to build a system that minimises the difference between these two corresponding features. In order to effectively model the properties of the HVS, a couple of related systems have been proposed. Zhang et al. [51] proposed a similarity index metric which calculated the phase congruence and gradient magnitude to represent the HVS system, while [46] implemented an efficient standard deviation pooling strategy which demonstrated that the gradient magnitude of an image still holds true as a technique for representing the HVS. [51] adopted a novel deviation pooling technique to compute the quality score from the gradient and chromaticity similarities as a measure for local structural distortions.

The Banach Wasserstein GAN (BWGAN) is a framework that makes use of arbitrary norms other than the popularly used l_2 norm as the underlying metric of choice in adversarial training. Adler et al. in [0] translated the WGAN-GP model to Banach spaces which have the capacity to utilize norms that capture desired image features like edges, texture, etc.

3 Our Approach

We present a single GAN model capable of implementing image-to-image synthesis. We combine the benefits of established FR-IQA metrics [51, 53, 51] and the low-level salient features of a deep convolutional neural network (DCNN) to aid adversarial image synthesis aimed at producing perceptually appealing images. In our model, we introduced an attention schema that exploits the salient perceptual features in a channel-wise fashion and the spatial map representation embeddings of standard FR-IQA metrics (SSIM, MDSI and FSIM). Our framework consists of five main components; a quality-aware generator network $G : (g_1, g_2)$, where g_1 and g_2 represent the encoder and decoder section, respectively. g_1 and g_2 are coupled with a perceptual attentive mechanism (PAM) for quality encoding and a perceptual quality map fusion network M_F at the latent space as shown in Figure 1. The discriminative networks $D : (d_1, d_2)$ critics images generated by $G : (g_1, g_2)$ in an adversarial manner without compromising image quality and the perceptual consistency with the reference image. The perceptual quality map generator combines the core quality metric functions that capture the sensitive perceptual features of a given image, while the score regression network
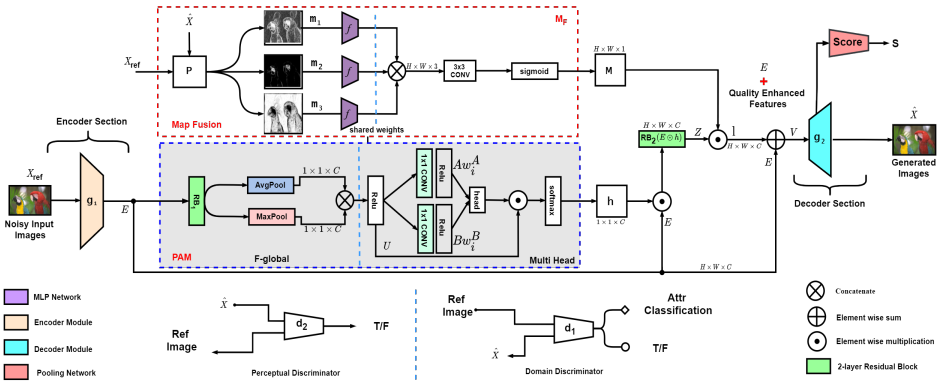
Figure 1: The quality encoding architecture. The structure shows the generator $G : (g_1, g_2)$ configuration coupled with a quality Map Fusion network M_F. The domain discriminator d_1 (lower-right) extracts features where True/False predictions are made per pixel and attribute classification is executed to ensure multi-domain adaptation, d_2 critics the perceptual map to maintain consistency in an adversarial manner.

pools the images synthesized by $G : (g_1, g_2)$ to estimate reference quality scores s. Our overall objective consists of a Wasserstein Gradient Penalty, Structural Similarity Index Gradient Penalty (SSIM-GP) and a Natural Image Quality Estimator (NIQE) as defined in section 4.

3.1 Perceptual Attention Mechanism (PAM)

The attention mechanism augments perceptual features from a prior generator encoder network $g_1(\cdot)$ computed over input images X_i. The aim is to establish a convex combination of quality-enhanced condensed representations of the input image for real time training. We begin by describing the channel attention in PAM, which is based on the CBAM module [45]. PAM involves two steps: first, per-channel "summary statistics" F_{global} obtained from a 2-layer residual block RB_1, is calculated to yield the global feature attention vector $U \in \mathbb{R}^{1 \times 1 \times C}$. Secondly, a multi-head network $head_i = Attention(Aw_i^A, Bw_i^B)$ applies a non-linear multi-head attention transformation which allows the model to jointly attend to information A and B from different representation sub-spaces w_i^A and w_i^B of the bottleneck [41]. The channel-based attention output given as $h = \text{softmax}(head_i \odot U)$ is multiplied with the encoder output $E \in \mathbb{R}^{H \times W \times C}$ from g_1 and processed by residual block RB_2 to produce channel-based attention embeddings, denoted as $Z = RB_2(E \odot h)$ where $Z \in \mathbb{R}^{H \times W \times C}$.

3.2 Perceptual Map Generator

We selected the FSIM [51], MS-SSIM [33], and MDSI [51] image quality metrics to generate similarity maps because the trio collectively capture key image characteristics that are similar to the HVS [3, 31, 33, 44, 51] as shown in Figure 2. The FSIM metric captures the luminance, contrast and structural information. For the MS-SSIM metric, we considered multiple scales of the synthesised image and its reference for contrast and structure while the MDSI map is derived by extracting the gradient and chromaticity of the pair of synthesised and reference images, respectively. We use an intensity coefficient; $0.3 \leq \alpha \leq 1$ to specify the intensity of the maps. The map fusion network M_F is divided into three stages. First, we extract the feature similarity representations between the reference image X_{ref} and the generated images $\hat{X}_i$ per iteration given as $p(X_{ref}, \hat{X}_i)$ for the aforementioned similarity metrics, where
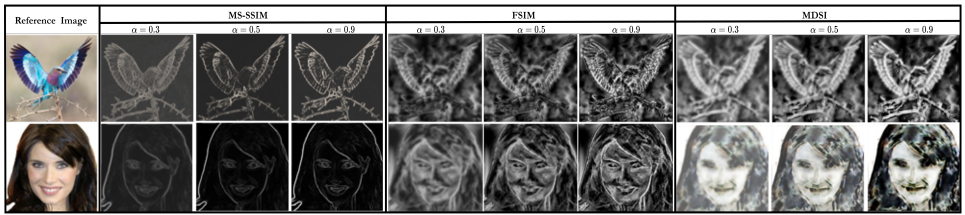
Figure 2: A spectrum of quality maps obtained at different intensities.

$p(\cdot)$ is an arbitrary function used to calculate similarity index maps; MS-SSIM, FSIM and MDSI. Secondly, the generated maps, $(m_1, m_2,$ and $m_3)$ are concatenated and pre-processed by two-layer MLP networks $f(\cdot)$ to form a spatial-based perceptual map representation $M \in \mathbb{R}^{H \times W \times 1}$. At the last stage, the predicted future states I is then computed as the expectation of spatial features M and the channel-based features $Z \in \mathbb{R}^{H \times W \times C}$. I is then summed with the output of the encoder E given as $V = I + E$. The resulting output $V \in \mathbb{R}^{H \times W \times C}$ which is fed to decoder g_2 represents latent features that are optimized for better image quality.

4 Banach space gradient penalty

Quality assessment metrics for the distance between images has been limited to cost functions that take the form of l_1 or l_2 norms. However, issues like non-convexity and complications in gradient computations (vanishing gradients, exploding gradients, etc) are some of the struggles experienced in formulating optimization problems. To mitigate the aforementioned computational shortcomings, the Wasserstein distance was introduced in [2].

However, a wide variety of untapped metrics [53, 44] exist that can be used to compare and emphasize key features of interest. In this regard, we extended the Wasserstein distance beyond the popular WGAN with the gradient penalty (WGAN-GP), which is constrained to l_2 norms and rather adapted a more complete space called the Banach space [1]. Our technique, similar to [1, 4, 14] shows that the characterisation of γ-lipschitz functions via the norm of the differential can be extended from the l_2 setting to arbitrary Banach spaces by considering the gradient as an element in the dual of B. Such a loss function is given as:

$$L_B = \lambda E_Y \left(\frac{1}{\gamma} \|\partial D(Y)\|_B^* - 1 \right)^2, \tag{1}$$

where $\lambda, \gamma \in \mathbb{R}$, are regularization parameters. These Banach space norms give room for specific image features such as texture, structure, contrast and luminance which highlight the perceptual appeal of a human observer, as described in section 4.1 and 4.2.

4.1 Structural Similarity (SSIM) index

The SSIM index measures the perceptual difference between two similar images. The local mean, variance and structure are computed to find an local quality score [44]. The SSIM index computes changes to local mean, local variance and local structure between two images X and $\hat{X}$. The local scores are then averaged across the image to find the image quality score.

$$L(X_{(i,j)}, \hat{X}_{(i,j)}) = \frac{2\mu_x(i,j)\mu_{\hat{x}}(i,j) + C_1}{\mu_x^2(i,j) + \mu_{\hat{x}}^2(i,j) + C_1}, \ CS(X_{(i,j)}, \hat{X}_{(i,j)}) = \frac{2\sigma_{x\hat{x}}(i,j) + C_2}{\sigma_{x^2}(i,j) + \sigma_{\hat{x}^2}(i,j) + C_2}, \tag{2}$$

$$d_b(X_{(i,j)}, \hat{X}_{(i,j)}) = \sqrt{2 - L((X_{(i,j)}, \hat{X}_{(i,j)})) - CS(X_{(i,j)}, \hat{X}_{(i,j)})}, \tag{3}$$

Figure 3: A sample of synthesised images representing different datasets.

where X and $\hat{X}$ refer to the input and synthesised images, the subscript (i,j) is the pixel index, $\mu_{(i,j)}$ and $\sigma_{(i,j)}$ are the local mean and standard deviation, respectively. $L(X_{(i,j)},\hat{X}_{(i,j)})$, $C(X_{(i,j)},\hat{X}_{(i,j)})$ and $S(X_{(i,j)},\hat{X}_{(i,j)})$ are the local luminance, contrast and structure scores at pixel (i,j), respectively. Furthermore, since $d_b(X,\hat{X})$ is bounded, the lipschitz constant can be imposed directly by introducing a gradient penalty regularization term given as:

$$SSIMGP = E_{X \sim P_X, \hat{X} \sim P_{\hat{X}}} \left[\left(\frac{|D(X) - D(\hat{X})|}{d_b(X,\hat{X})} \right) - 1 \right]^2. \tag{4}$$

This makes the SSIM a good candidate for quality awareness which is beneficial for regularizing GANs. The complete mathematical properties are described in [6].

4.2 Natural Image Quality Estimator (NIQE)

The NIQE [26] is an NR-IQA metric of perceptually relevant spatial domain Natural Scene Statistics (NSS) features extracted from local image patches that capture the essential low-order statistics of natural images. The equation is given as:

$$\hat{I}(i,j) = \frac{I_{(i,j)} - \mu_{(i,j)}}{\sigma_{(i,j)} + 1}, \tag{5}$$

where $I_{(i,j)}$ is the pixel index and $\mu_{(i,j)}$ and $\sigma_{(i,j)}$ are the local mean and standard deviation. The NIQE captures the naturalness of a pristine reference image by modelling a generalized gaussian distribution (GGD) [57], and models the products of neighbouring pixel coefficients using an Asymmetric GGD (AGGD). The parameters of both the GGD and AGGD are then modelled using a Multivariate Gaussian Model (MVG) distribution [28]. The quality of the test image is measured in terms of the "distance" of its MVG parameters μ_t and σ_t from the pristine MVG parameters obtained. Finally, discriminator gradients computed for both pristine reference and synthesised images are used to compute the distance between the pair. The expression is given as:

$$\|(\mu_X, \Sigma_X)\|_{NIQE} := \sqrt{(\mu_X - \mu_{\hat{X}})^T \left(\frac{\Sigma_X + \Sigma_{\hat{X}}}{2} \right)^{-1} (\mu_X - \mu_{\hat{X}})}, \tag{6}$$

where μ_X, $\mu_{\hat{X}}$, Σ_X and $\Sigma_{\hat{X}}$ are the mean and covariance of the reference X and synthesised $\hat{X}$ images, respectively. In addition to the SSIM and NIQE metrics, we also used a 1-GP regularizer [9] designed to force the local statistics of the discriminator gradient to be as close to those of real images. Our claim is that such a regularization strategy results in improving visual quality of the generated images especially for attributes like hair, age, skin
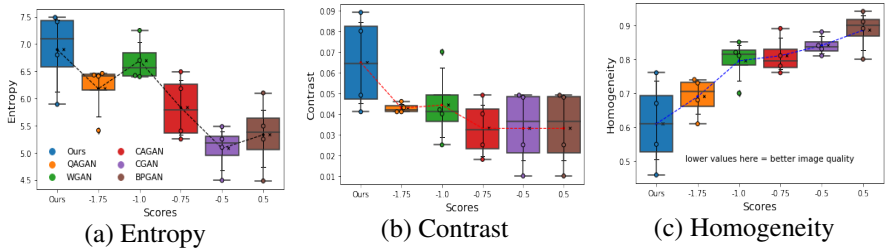
(a) Entropy   (b) Contrast   (c) Homogeneity

Figure 4: Statistical feature IQA metric values.

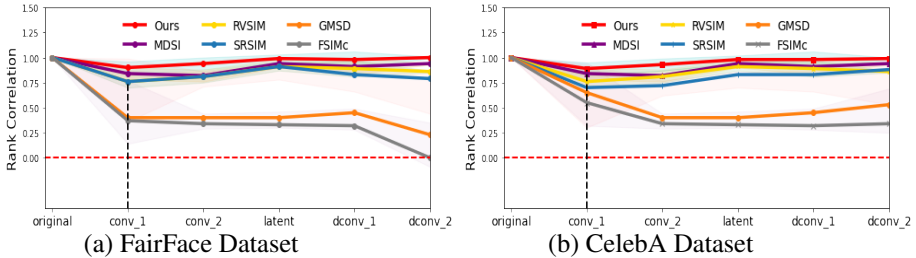
(a) FairFace Dataset   (b) CelebA Dataset

Figure 5: Superman's rank correlation values at different layers of the network

colour etc. We worked in the WGAN-GP framework to demonstrate our method. The overall discriminator cost function includes the NIQE function regularizer, the SSIM and the 1-GP regularizer defined as:

$$L_{BP} = \lambda_1 E_{\hat{x} \sim P_{\hat{x}}}(\left\| \nabla_{\hat{x}} D(\hat{X}) | \mu_P, \Sigma_P \right\|_{NIQE}) + \lambda_2 E_{x \sim P_x, \hat{x} \sim P_{\hat{x}}} \left[\left(\frac{|D(X) - D(\hat{X})|}{d_b(X, \hat{X})} \right) - 1 \right]^2_{SSIM}$$

$$+ \lambda_3 E_{x \sim P_{\hat{x}}}(\left\| \nabla_{\hat{x}} D(\hat{X}) | \right\|_{1-GP}),$$

The full objective is given as: (7)

$$L_{GAN}(G, D, X, \hat{X}) = E[logD(X)) + E[log(1 - D(G(\hat{X})) - s)] + L_{BP}, \quad (8)$$

where s is the generated score from the regression network minimised over the groudtruth scores of the images. we use λ_1, λ_2 and λ_3 as a means of tuning the objective functions to achieve better results.

5   Training Strategy

We trained our model using the Adam optimizer, with momentum values set at $\beta_1 = 0.5$ and $\beta_2 = 0.99$, we used a batch size of 8 for most experiments on CelebA [24], Celeba-HQ [22] and FairFAce [16], respectively. A learning rate of 0.0001 for the first 10 epochs which linearly decayed to 0 over the corresponding epochs. We trained the entire model on three NVIDIA Titan X GPUs.

5.1   Datasets

We evaluated the efficacy of our proposed technique on the following datasets: The Celeb-Faces Attributes (CelebA) [24] of 202,599 celebrity face images. We cropped the initial images to 178x178, then resized them to 64x64. The CIFAR-10 [21] dataset consists of 60,000 32x32 colour images in 10 classes, with 6,000 images per class. The Fair-Face [17] image dataset contains 108,501 images, with an emphasis of balanced race composition in the dataset comprising 7 race groups: White, African, Indian, East Asian, Southeast Asian,
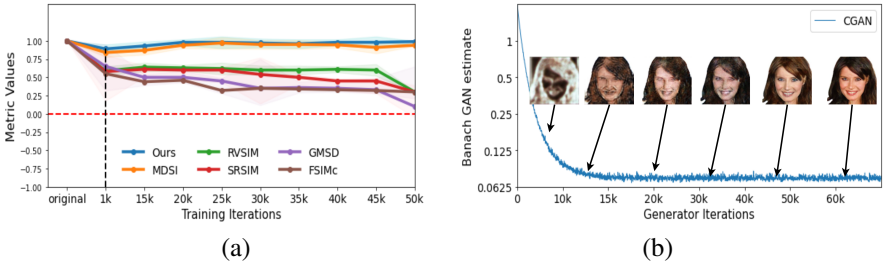
Figure 6: HOG similarity metrics (left) and synthesised image results for FairFace Dataset.
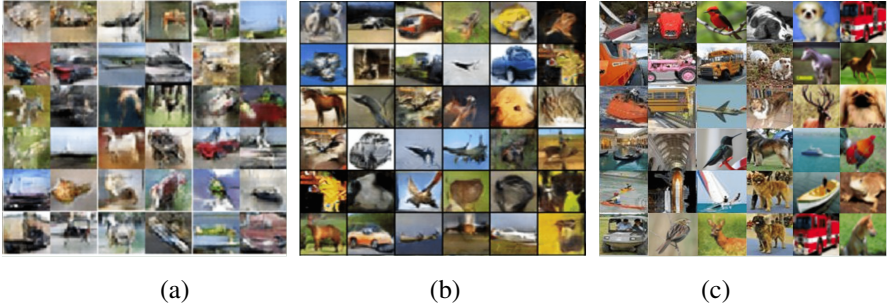

(a)　　　　　　　　　(b)　　　　　　　　　(c)

Figure 7: Randomly sampled images generated using a combination of our model base line (*BL*) and different losses for W-GAN and BWGAN (SSIM and NIQE) of the CIFAR-10 dataset. 7 (a) Shows images synthesised using *BL* and just the WGAN with gradient penalty loss. 7 (b) performs better when the SSIM loss function is added (*BL+WGAN+SSIM*). 7 (c) shows the best results when all loss functions are included (*BL+WGAN+SSIM+NIQE*).

Middle East, and Hispanic. For evaluations, we used the LIVE [59] which consists of 982 distorted images with 5 different distortions. The TID2008 dataset [56] that contains 25 reference images and a total of 1,700 distorted images. We also used Edges-to-shoes 50,000 training images from UT Zappos50K dataset [49] and Edges-to-Handbag 137,000 Amazon Handbag images from [55], trained for 15 epochs and batch size 8.

5.2 Evaluation

To evaluate the performance of the synthesized images as shown in Figure 3, two key evaluation criteria were adopted in our paper; the Spearman's Rank Order Correlation Coefficient (SROCC) and the Linear Correlation Coefficient (LCC) [7]. SROCC is a measure of the monotonic relationship between the ground-truth and model prediction, while the LCC is a measure of the linear correlation between the ground-truth and model prediction. Table 1 and 2 shows the SROCC and LCC performance of the competing IQA methods for different distortion types, respectively. In general, our model performs competitively among most distortion types. Compared with BPSQM, our model shows more performance of about 4.5% overall in dealing with the distortion of AGN, SCN, HFN, JPEG and MN, respectively as indicated in Table 1. For comparison with previous models, we computed three quantitative measures: Inception Score (IS), Frechet Inception Distance (FID) and the Feature Similarity (FSIM) index. IS measures the sample quality and diversity by finding the entropy of the predicted labels. FID score measures the similarity between real and fake samples by fitting a multivariate Gaussian (MVG) model to the intermediate representation. The FSIM index computes quality estimates based on phase congruency as the primary feature, and incor-

(a)   (b)   (c)
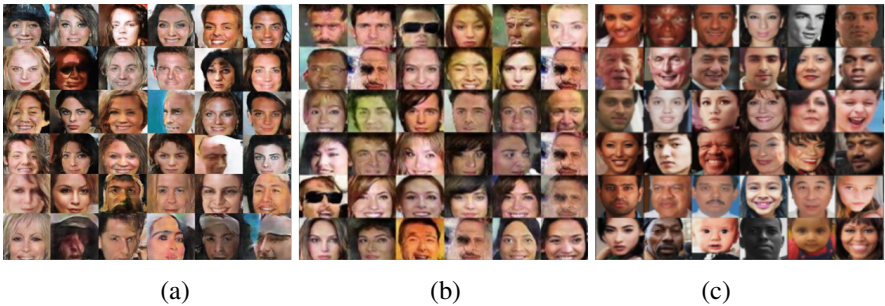
Figure 8: Randomly sampled images generated using different models for FairFace dataset. 8 (a) Shows images synthesised using DCGAN. 8 (b) performs better due to the gradient penalty approach of WGAN. However, 8 (c) shows that our model performs even better when Banach losses are included.

Table 1: SROCC comparison on individual distortion types on the TID2008 databases.

SROCC	TID2008												
	AGN	ANMC	SCN	MN	HFN	IMN	QN	GB	DEN	JPEG	JP2K	JGTE	J2TE
GMSD [46]	0.911	**0.888**	0.914	0.747	0.919	0.683	0.857	0.911	**0.966**	0.954	0.983	0.852	0.873
FSIMc [51]	0.910	0.864	0.890	0.863	0.921	0.736	0.865	0.949	0.964	0.945	0.977	0.878	**0.884**
BLIINDSII [31]	0.779	0.807	0.887	0.691	0.917	0.908	0.851	0.952	0.908	0.928	0.940	0.865	0.855
DIIVINE [29]	0.812	0.844	0.854	0.713	0.922	0.915	0.874	0.943	0.912	0.930	0.938	0.873	0.852
BRISQUE [25]	0.853	0.861	0.885	0.810	0.931	**0.927**	0.881	0.933	0.924	0.934	0.944	**0.891**	0.836
NIQE [24]	0.786	0.832	0.903	0.835	0.931	0.913	0.893	**0.953**	0.917	0.943	0.956	0.862	0.827
BPSQM [50]	0.881	0.801	0.935	0.786	0.938	**0.933**	0.920	0.937	0.914	0.943	0.967	0.829	0.644
Ours	**0.936**	0.878	**0.961**	**0.939**	**0.948**	0.892	0.915	0.898	0.878	**0.955**	**0.987**	0.836	0.779

porates the gradient magnitude as the complementary feature for the real and fake samples, respectively. Table 3 shows the quantitative comparison of the GAN-metric performance for BPGAN [43], CAGAN [50], CGAN [56] , WAGAN [4], QAGAN [14] and ours for CelebA dataset. We also carried out pixel variation analysis on the synthesised images by using the second order features of the synthesized images, which are based on the gray level co-occurrence matrix (GLCM) [11]. We used the aforementioned technique to determine the Entropy, Homogeneity and Correlation of the synthesised images in comparison with state-of-the-art methods as shown in Figure 4. Entropy is useful for assessing sharpness while Homogeneity and Correlation are useful for evaluating the Contrast of an image. Entropy and Correlation increase in image quality, whereas Homogeneity energy values decrease with increase in image quality. From the Entropy plot in 5(a), our model performs decently well by over 3.5% compared to the QAGAN and WAGAN. The Contrast level improves drastically for our approach as compared to the other methods that are closely matched at a tolerance of about 2%. We also observed that most models possess similar homogeneity values except our model and QAGAN which reflect significant performance values.

5.3   Ablation Study

Ablation studies on our loss functions was implemented to test model robustness in general for the CIFAR-10, FairFace and CelebA datasets, respectively. The Lagrange coefficients λ_1 and λ_2 of the SSIM and NIQE losses were also changed empirically within ($0.001 \geq \lambda_1$ and $\lambda_{21} \leq 1.000$) range, to check the effect on the perceptual appeal of the synthesised images. It was inferred that reducing the coefficients towards the lower limit weakens the discriminative power which in turn reduces the quality of the synthesised images from the generator. We

Table 2: LCC evaluation on LIVE database.

LCC	LIVE					
	JP2K	JPEG	WN	BLUR	FF	ALL
BRISQUE [■]	0.923	0.973	0.985	0.951	0.903	0.942
CORNIA [■]	0.951	0.965	0.987	0.968	0.917	0.935
CNN [■]	0.953	0.981	0.984	0.953	0.933	0.953
SOM [■]	0.952	0.961	0.991	0.974	0.954	0.962
BIECON [■]	0.965	**0.987**	0.970	0.945	0.931	0.962
Ours	**0.975**	0.986	**0.994**	**0.988**	**0.960**	**0.982**

Table 3: GAN-metric performance.

Model	CelebA		
	FID↓	FSIM↑	IS↑
BPGAN [■]	86.10	69.13	0.87
CGAN [■]	43.21	71.10	0.89
CAGAN [■]	36.16	71.33	0.90
WGAN [■]	33.24	72.60	0.91
QAGAN [■]	18.23	82.69	0.96
Ours	**18.39**	**83.40**	**0.97**

Table 4: Ablation study on CelebA and CIFAR-10 datasets on our model baseline "BL" with a combination of the quality modules "PAM" and "M_F" and losses "SSIM" and "NIQE".

BL	PAM	M_F	NIQE	SSIM	CIFAR-10		CelebA	
					FID↓	IS↑	FID↓	IS↑
✓					38.10 ± 0.12	8.20 ± 0.03	29.80 ± 0.10	0.86 ± 0.03
✓	✓			✓	19.01 ± 0.10	8.01 ± 0.13	13.16 ± 0.02	0.88 ± 0.19
✓	✓		✓		16.31 ± 0.21	8.00 ± 0.35	11.86 ± 0.07	0.89 ± 0.11
✓		✓		✓	15.00 ± 0.20	**7.46 ± 0.21**	10.76 ± 0.43	0.90 ± 0.10
✓		✓	✓		13.21 ± 0.10	7.80 ± 0.10	**6.38 ± 0.39**	0.96 ± 0.10
✓	✓	✓	✓	✓	**8.06 ± 0.22**	7.48 ± 0.62	6.40 ± 0.71	**0.97 ± 0.16**

also conducted a Histogram of Oriented Gradient (HOG) similarity performance with the Inception v3 model [40] for the input and synthesised images on the FairFace dataset, in order to obtain the model layer-wise performance at specific iterations of the baseline of our model. Figure 6 (a) shows the HOG similarity performance at different iterations while training for our model compared to other quality metric techniques.

Our results show that our approach is closest to MDSI [51], as compared to RVSIM [43], GSMD [46], SRSIM [52], FSIMc [51] that perform slightly below our model. An SRCC plot representation in Figure 5 depicts the rank correlation performance for both FairFace and CelebA dataset. The values confirm that our model performs favourably over other aforementioned techniques. At different iteration values, we also observed decent image quality improvements at about 20k - 30k iterations as shown in Figure 6(b) for the FairFace dataset. Figures 7 and 8 show further results obtained from a combination of different loss functions and other competitive models, respectively.

We computed the FID and IS scores of synthesised images for ClebeA and CIFAR-10 datasets with resolutions of 64 x 64 and 32 x 32, respectively. Table 4 shows the performance of our model baseline (BL) for different combinations of attention schemes (PAM and M_F) and the IQA losses (NIQE and SSIM). By observation, we see from Table 4 that including the M_F module significantly boost image quality, this is a confirmation that perceptual spatial salient maps are crucial in GAN models for better image quality [13, 38, 47].

Furthermore, we applied the PAM and M_F attention modules to the StleGAN2 [19] architecture. We also added the proposed Banach space norms (SSIM NIQE) to compare the overall model performance with our model. In Table 5, we show the trade-offs of the QA-GAN [14], StyleGAN2 [19] and our model baseline (BL). We used different combinations of the standard IQA metrics as discussed in section 4.1 and 4.2. Our findings confirm that our approach is competitive with state-of-the-art. Most importantly, we see improved performance of our model at lower resolutions (32 x 32), this improvement can be attributed to the attention schema employed. In Figure 9, we showcase the performance of QAGAN [14] and our model on image synthesis for CelebA dataset. Our results show that our model performs significantly well overall, Table 5 gives a clearer representation of the performance levels.

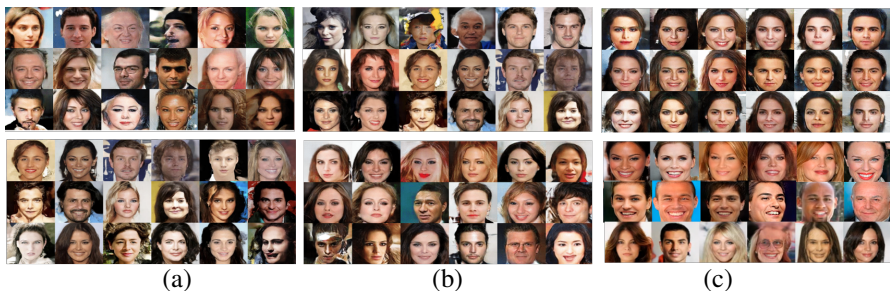| (a) | (b) | (c) |

Figure 9: Randomly sampled images for QAGAN [14] (Top row) and our model (Bottom row) with different losses (SSIM and NIQE) of the CelebA dataset. 10 (a) Shows images synthesised using the baseline of both models. 9 (b) The SSIM loss function is added for both cases. We observe similar performance levels. 9 (c) Our model performs better overall with the best results when all loss functions are included ($SSIM + NIQE$).

Table 5: FID on CelebA and CIFAR-10 dataset.

Model	CIFAR-10 (32 x 32) FID↓	CelebA (64 x 64) FID↓
QAGAN [14]	41.20 ± 0.25	10.03 ± 0.35
QAGAN (SSIM) [14]	14.13 ± 0.32	6.44 ± 0.43
QAGAN (NIQE) [14]	12.57 ± 0.11	6.40 ± 0.23
QAGAN (SSIM + NIQE) [14]	10.01 ± 0.13	$\mathbf{6.16 \pm 0.05}$
StyleGAN2 [19]	37.11 ± 0.15	9.03 ± 0.25
StyleGAN2 (SSIM) [19]	13.14 ± 0.02	$\mathbf{5.84 \pm 0.13}$
StyleGAN2 (NIQE) [19]	11.17 ± 0.31	6.10 ± 011
StyleGAN2 (SSIM + NIQE) [19]	10.81 ± 0.13	6.18 ± 0.05
BL	37.80 ± 0.22	10.06 ± 0.43
BL (SSIM)	$\mathbf{12.80 \pm 0.12}$	6.86 ± 0.62
BL (NIQE)	10.20 ± 0.79	6.36 ± 0.44
BL (SSIM + NIQE)	$\mathbf{9.76 \pm 0.37}$	6.21 ± 0.36

6   Conclusion

In this paper, we introduced a novel quality encoding protocol that harnesses the image quality maps mimicking the HVS and the perceptual properties from a deep convolutional neural network (DCNN) to provide perceptually consistent features that translate to better image quality. We identified visually sensitive parameters and adapted a quality perceptual attention scheme that narrows down these features to a localised embedding which incentives perceptual representations over other features. The aim was to target the most relevant intrinsic features responsible for image texture, structural contrast and luminance which we use to guide the adversarial model towards high quality image synthesis. We also introduced a critic model that monitors perceptual consistency for each image representation. We demonstrated state-of-the-art or comparable performance over other approaches.

References

[1] J. Adler and S. Lunz. Banach wasserstein GAN. In *NeurIPS*, 2018.

[2] Poorya Aghdaie, Baaria Chaudhary, Sobhan Soleymani, Jeremy M. Dawson, and Nasser M. Nasrabadi. Attention aware wavelet-based detection of morphed face im-

ages. *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8, 2021.

[3] Poorya Aghdaie, Baaria Chaudhary, Sobhan Soleymani, Jeremy M. Dawson, and Nasser M. Nasrabadi. Detection of morphed face images using discriminative wavelet sub-bands. *2021 IEEE International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, 2021.

[4] Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *ArXiv*, abs/1701.07875, 2017.

[5] Dominique Brunet, E. Vrscay, and Z. Wang. On the mathematical properties of the structural similarity index. *IEEE Transactions on Image Processing*, 21:1488–1499, 2012.

[6] Scott J. Daly. Visible differences predictor: an algorithm for the assessment of image fidelity. In Bernice E. Rogowitz, editor, *Human Vision, Visual Processing, and Digital Display III*, volume 1666, pages 2 – 15. International Society for Optics and Photonics, SPIE, 1992.

[7] Ron N. Forthofer and Robert G. Lehnen. *Rank Correlation Methods*, pages 146–163. Springer US, Boston, MA, 1981.

[8] I. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, B. Xu, David Warde-Farley, S. Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. *ArXiv*, abs/1406.2661, 2014.

[9] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein ǵans. In *NIPS*, 2017.

[10] Y. Han, Yunze Cai, Yin Cao, and Xiaoming Xu. A new image fusion performance metric based on visual information fidelity. *Inf. Fusion*, 14:127–135, 2013.

[11] R. Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67:786–804, 1979.

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2017.

[13] J. Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. *ArXiv*, abs/1603.08155, 2016.

[14] Parimala Kancharla and Sumohana S. Channappayya. Quality aware generative adversarial networks. In *NeurIPS*, 2019.

[15] Le Kang, Peng Ye, Yi Li, and David S. Doermann. Convolutional neural networks for no-reference image quality assessment. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1733–1740, 2014.

[16] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *ArXiv*, abs/1908.04913, 2019.

[17] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1547–1557, 2021.

[18] Tero Karras, Timo Aila, S. Laine, and J. Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. *ArXiv*, abs/1710.10196, 2018.

[19] Tero Karras, S. Laine, M. Aittala, Janne Hellsten, J. Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8107–8116, 2020.

[20] Jongyoo Kim and S. Lee. Fully deep blind image quality predictor. *IEEE Journal of Selected Topics in Signal Processing*, 11:206–220, 2017.

[21] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[22] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. *ArXiv*, abs/1907.11922.

[23] Yudong Liang, J. Wang, Xingyu Wan, Y. Gong, and N. Zheng. Image quality assessment using similar scene as reference. In *ECCV*, 2016.

[24] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[25] Anish Mittal, Anush K. Moorthy, and Alan C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21:4695–4708, 2012.

[26] Anish Mittal, R. Soundararajan, and A. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20:209–212, 2013.

[27] Salman Mohamadi and Hamidreza Amindavar. Deep bayesian active learning, a brief survey on recent advances. *ArXiv*, abs/2012.08044, 2020.

[28] Anush K. Moorthy and A. Bovik. Statistics of natural image distortions. *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 962–965, 2010.

[29] Anush K. Moorthy and Alan C. Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing*, 20:3350–3364, 2011.

[30] Moktari Mostofa, Syeda Nyma Ferdous, Benjamin S. Riggan, and Nasser M. Nasrabadi. Joint-srvdnet: Joint super resolution and vehicle detection network. *IEEE Access*, 8:82306–82319, 2020.

[31] Hossein Ziaei Nafchi, Atena Shahkolaei, Rachid Hedjam, and Mohamed Cheriet. Mean deviation similarity index: Efficient and reliable full-reference image quality evaluator. *IEEE Access*, 4:5579–5590, 2016.

[32] Uche M. Osahor, Hadi Kazemi, Ali Dabouei, and Nasser M. Nasrabadi. Quality guided sketch-to-photo image synthesis. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3575–3584, 2020.

[33] Alexander Osokin and Dmitry Sidorov. Modification of ssim metrics. In Alexander Dudin, Anatoly Nazarov, Rafael Yakupov, and Alexander Gortsev, editors, *Information Technologies and Mathematical Modelling*, pages 351–355, Cham, 2014. Springer International Publishing.

[34] Da Pan, Ping Shi, Ming Hou, Zefeng Ying, Sizhe Fu, and Yuan Zhang. Blind predicting similar quality map for image quality assessment. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6373–6382, 2018.

[35] Soo-Chang Pei and Li-Heng Chen. Image quality assessment using human visual dog model fused with random forest. *IEEE Transactions on Image Processing*, 24:3282–3292, 2015.

[36] Rafael Reisenhofer, Sebastian Bosse, Gitta Kutyniok, and Thomas Wiegand. A haar wavelet-based perceptual similarity index for image quality assessment. *Signal Process. Image Commun.*, 61:33–43, 2018.

[37] D. Ruderman. The statistics of natural images. *Network: Computation In Neural Systems*, 5:517–548, 1994.

[38] Michele A. Saad, Alan C. Bovik, and Christophe Charrier. Blind image quality assessment: A natural scene statistics approach in the dct domain. *IEEE Transactions on Image Processing*, 21:3339–3352, 2012.

[39] H. Sheikh, M. Sabir, and A. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15:3440–3451, 2006.

[40] Christian Szegedy, V. Vanhoucke, S. Ioffe, Jonathon Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

[41] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *ArXiv*, abs/1706.03762, 2017.

[42] Fulai Wang, Xian Sun, Zhi Guo, Yu Huang, and Kun Fu. An object-distortion based image quality similarity. *IEEE Signal Processing Letters*, 22:1534–1537, 2015.

[43] Nannan Wang, Wenjin Zha, Jie Li, and Xinbo Gao. Back projection: An effective post-processing method for gan-based face sketch synthesis. *Pattern Recognit. Lett.*, 107:59–65, 2017.

[44] Zhou Wang, A. Bovik, H. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612, 2004.

[45] S. Woo, Jongchan Park, Joon-Young Lee, and In-So Kweon. Cbam: Convolutional block attention module. In *ECCV*, 2018.

[46] Wufeng Xue, Lei Zhang, X. Mou, and A. Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Transactions on Image Processing*, 23:684–695, 2014.

[47] Xiaohan Yang, Fan Li, Wei Zhang, and Lijun He. Blind image quality assessment of natural scenes based on entropy differences in the dct domain. *Entropy*, 20:885, 2018.

[48] Peng Ye, Jayant Kumar, Le Kang, and David S. Doermann. Unsupervised feature learning framework for no-reference image quality assessment. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1098–1105, 2012.

[49] A. Yu and K. Grauman. Fine-grained visual comparisons with local learning. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 192–199, 2014.

[50] Jun Yu, Xingxin Xu, Fei Gao, Shengjie Shi, Meng Wang, Dacheng Tao, and Qingming Huang. Towards realistic face photo-sketch synthesis via composition-aided GANs. volume 51, pages 4350–4362, 2021.

[51] L. Zhang, Lei Zhang, X. Mou, and D. Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20:2378–2386, 2011.

[52] Lin Zhang and Hongyu Li. Sr-sim: A fast and high performance iqa index based on spectral residual. *2012 19th IEEE International Conference on Image Processing*, pages 1473–1476, 2012.

[53] Lin Zhang, Ying Shen, and Hongyu Li. Vsi: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 23: 4270–4281, 2014.

[54] P. Zhang, W. Zhou, L. Wu, and H. Li. SOM: Semantic obviousness metric for image quality assessment. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2394–2402, 2015.

[55] Jun-Yan Zhu, Philipp Krähenbühl, E. Shechtman, and Alexei A. Efros. Generative visual manipulation on the natural image manifold. In *ECCV*, 2016.

[56] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.