GIQA: Generated Image Quality Assessment

Shuyang Gu[1], Jianmin Bao[2⋆], Dong Chen[2], and Fang Wen[2]

[1] University of Science and Technology of China
[2] Microsoft Research
gsy777@mail.ustc.edu.cn and {jianbao, doch, fangwen}@microsoft.com

Abstract. Generative adversarial networks (GANs) achieve impressive results today, but not all generated images are perfect. A number of quantitative criteria have recently emerged for generative models, but none of them are designed for a single generated image. In this paper, we propose a new research topic, Generated Image Quality Assessment (GIQA), which quantitatively evaluates the quality of each generated image. We introduce three GIQA algorithms from two perspectives: learning-based and data-based. We evaluate a number of images generated by various recent GAN models on different datasets and demonstrate that they are consistent with human assessments. Furthermore, GIQA is available for many applications, like separately evaluating the realism and diversity of generative models, and enabling online hard negative mining (OHEM) in the training of GANs to improve the results.

Keywords: generative model, generative adversarial networks, image quality assessment

1 Introduction

Recent studies have shown remarkable success in generative models for their wide applications like high quality image generation [18, 2], image-to-image translation [15, 35, 11, 10], data augmentation [7, 6], and so on. However, due to the large variance in quality of generated images, not all generated images are satisfactory for real-world applications. Relying on a manual quality assessment of generated images takes a lot of time and effort. This work proposes a new research topic: Generated Image Quality Assessment (GIQA). The goal of GIQA is to automatically and objectively assess the quality of each image generated by the various generative models.

GIQA is related to Blind/No-Reference Image Quality Assessment (NR-IQA) [38, 8, 30, 25, 34]. However, NR-IQA mainly focuses on quality assessment of natural images instead of the generated image. Most of them are distortion-specific; they are capable of performing NR-IQA only if the distortion that afflicts the image is known beforehand, *e.g.*, blur or noise or compression and so on. While the generated images may contain many uncertain model specific artifacts like checkboards [26], droplet-like [20], and unreasonable structure [40]. Unlike

⋆ Corresponding author. Please refer to our arxiv version for more paper details.

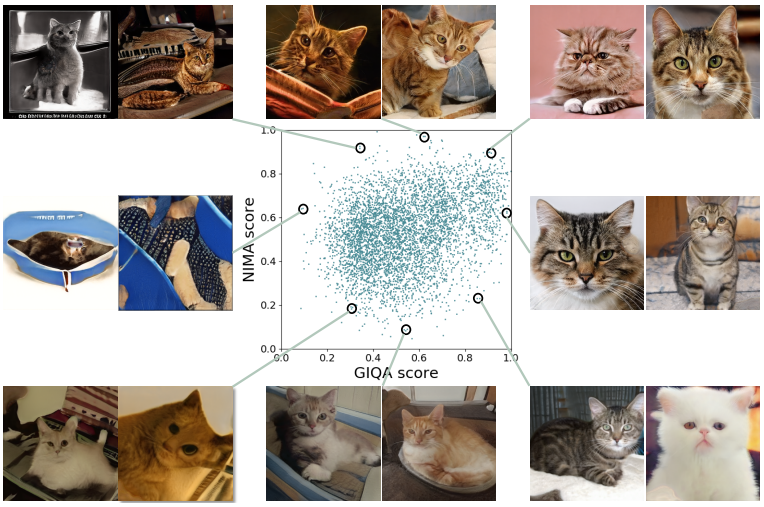

Fig. 1: Score distribution of a NR-IQA method NIMA [34] and our GIQA methods. Score of NIMA is normalized to [0,1] for better comparison, higher score denotes higher image quality. Our GIQA score is more consistent with human observation.

low-level degradations, these artifacts are difficult to simulate at different levels for training. Therefore, traditional natural image quality assessment methods are not suitable for generated images as shown in Figure 1. On the other hand, previous quantitative metrics, like Inception Score [12] and FID [13], focus on the assessment of the generative models, which also cannot be used to assess the quality of each single generated image.

In this paper, we introduce three GIQA algorithms from two perspectives: learning-based and data-based perspectives. For the learning-based method, we apply a CNN model to regress the quality score of a generated image. The difficulty is that it is hard to obtain the labelled quality score for a generated image. To address this problem, we propose a novel semi-supervised learning procedure. We observe that the quality of the generated images gets better and better during the training process of generative models. Based on this, we use images generated by models with different iterations, and use the number of iterations as the pseudo label of the quality score. To eliminate the label noise, we propose a new algorithm that uses multiple binary classifiers as a regressor to implement regression. Our learning-based algorithm can be applied to a variety of different models and databases without any manual annotation.

For data-based methods, the essence is that the similarity between the generated image and the real image could indicate quality, so we convert the GIQA problem into density estimation problem of real images. This problem can be broadly categorized as a parametric and non-parametric method. For the parametric method, we directly adopt the Gaussian Mixture Model (GMM) to capture the probability distribution of real data, then we estimate the probability

of a generated image as the quality score. Although this model is very simple, we find it works quite well for most situations. A limitation of the parametric method is that the chosen density might not capture complex distribution, so we propose another non-parametric method by computing the distance between generated image and its K nearest neighbours (KNN), the smaller distance indicates larger probability.

The learning-based method and the data-based method each have their own advantages and disadvantages. The GMM based method is easy to use and can be trained without any generated images, but it can only be applied to relatively simple distributed databases. The KNN based method has a great merit that there is no training phase, but its memory cost is large since it requires the whole training set to be stored. The learning-based method can handle a variety of complex data distributions, but it is also very time-consuming to collect the images generated by various models at different iterations. Considering both effectiveness and efficiency, we recommend GMM-GIQA mostly. We evaluate these 3 methods in detail in the experiments part.

The proposed GIQA methods can be applied in many applications. 1) We can apply it for generative model assessment. Current generative model assessment algorithms like Inception Score [12] and FID [13] evaluate the performance of the generative model in a score which represents the summation of two aspects: realism and diversity. Our proposed GIQA model can evaluate these two aspects separately. 2) By using our GIQA method, we can assess the quality of each generated image for a specific iteration of generator, and rank the quality of these samples, we suggest that the generator pay more attention to the samples with low quality. To achieve this, we adopt online hard negative mining (OHEM) [32] in the discriminator to put larger loss of weight to the lower quality generated samples. Thus the performance of the generator is improved by this strategy. 3) We can leverage GIQA as an image picker to obtain a subset of generated images with higher quality.

Evaluating the GIQA algorithm is an open and challenging problem. It is difficult to get the precise quality annotation for the generated images. In order to evaluate the performance of our methods, we propose a labeled generated image for the quality assessment (LGIQA) dataset. To be specific, we present a series of pairs which consist of two generated images for different observers to choose which has a better quality. We keep the pairs which are annotated with the consistent opinions for evaluating. We will release the data and encourage more research to explore the problem of GIQA.

To summarise, our main contribution are as follows, (1) To our knowledge, we are the first to propose the topic of GIQA. We proposed three straightforward methods from two perspectives to encourage further research. (2) GIQA is general and available to many applications, such as separately evaluating the quality and diversity of generative models and improving the results of generative model through OHEM. (3)We introduce the LGIQA dataset for evaluating different GIQA methods.

2 Related Work

In this section, we briefly review prior natural image quality assessment methods and generative model assessment methods that are most related to our work.

Image Quality Assessment: Traditional Image Quality Assessment(IQA) aims to assess the quality of natural images regarding low-level degradations like noise, blur, compression artifacts, *etc.* It is a traditional technique that is widely used in many applications. Formally, it can be divided into three main categories: Full-reference IQA (FR-IQA), Reduced-reference IQA (RR-IQA) and No-reference IQA (NR-IQA). FR-IQA is a relatively simple problem since we have the reference for the image to be assessed, the most widely used metrics are PSNR [14] and SSIM [36]. RR-IQA [37] aims to evaluate the perceptual quality of a distorted image through partial information of the corresponding reference image. NR-IQA is a more common real-world scenario which needs to estimate the quality of natural image without any reference images. Many NR-IQA approaches [38, 8, 30, 25] focus on some specific distortion. Recently advances in convolution neural networks (CNNs) have spawned many CNNs based methods [16, 17, 1] for natural image quality assessment. More recent works [28, 22] leverage the generative model in their framework to regress the final quality score.

Generative Model Assessment: Recent studies have shown remarkable success in generative models. Many generative models like VAEs [5], GANs [9], and Pixel CNNs [27] have been proposed, so the assessment of generative models has received extensive attention. Many works try to evaluate the generative model by conducting the user study, users are often required to score the generated images. While this will cost a large amount of time and effort. Therefore early work [31] propose a new metric Inception Score (IS) to measure the performance of generative model, the Inception Score evaluates the generative model in two aspects: realism and diversity of the generated images which are synthesized using the generative model. More recent work [13] proposes the Frchet Inception Distance (FID) score for the assessment of generative models. It takes the real data distribution into consideration and calculates the statistics between the generated samples distribution and real data distribution. [21] proposed precision and recall to measure generative model from quality and diversity separately.

3 Methods

Given a generated image $\mathcal{I}_g$, the target of GIQA is to quantitatively and objectively evaluate its quality score $S(\mathcal{I}_g)$ which should be consistent with human assessment. We propose solving this problem from two different perspectives. The first one is a learning-based method, in which we apply a CNN model to regress the quality score of a generated image. The second one is a data-based method, for which we directly model the probability distribution of real data. Thus we can estimate the quality of a generated image by the estimated probability from the model. We'll describe them in detail in the following sections.

3.1 Learning-based Methods

For learning-based methods, we aim to apply a CNN model to learn the quality of the generated images. Previous supervised learning method often require large amounts of labeled data for training. However, the quality annotation for the generated images is difficult to obtain since it is impossible for human observers to give the precise score to each generated image. Therefore, we propose a novel semi-supervised learning procedure.

Semi-Supervised learning: We find an important observation that the quality of generated images from most generative models, *e.g.*, PGGAN [18] and StyleGAN [19], is becoming better and better as the training iteration increases. Based on this, we collect images generated by models with different iterations, and use the number of iterations as the pseudo label of the quality score. Note that there is still a gap between the quality of the image generated by the last iteration and the real image. So we suppose that the quality of the generated images ranges from 0 to S_g, where $S_g \in (0, 1)$, and the quality of the real images is 1. Formally, the pseudo label of quality score $S_{\mathrm{p}}(\mathcal{I})$ for image $\mathcal{I}$ is

$$S_{\mathrm{p}}(\mathcal{I}) = \begin{cases} \frac{S_g \cdot iter}{max\_iter} & \textit{if } \mathcal{I} \textit{ is generated} \\ 1 & \textit{otherwise} \end{cases}, \tag{1}$$

where *iter* presents the iteration number, *max_iter* presents the maximum iteration number, S_g defines the maximum quality score for the generated image, we set it to 0.9 in our experiment. Then we are able to build a training dataset $\mathcal{D} = \{\mathcal{I}, S_{\mathrm{p}}(\mathcal{I})\}$ for semi-supervised learning, where $\mathcal{I}$ represents the generated images or the real images, $\mathcal{S}_{\mathrm{p}}(\mathcal{I})$ denotes the corresponding quality score.

Multiple Binary Classifiers as Regressor: A basic solution is to directly adopt a CNN based framework to regress the quality score from the input image. However, we find that this naive regression method is sub-optimal, since the pseudo label contains a lot of noise. Although statistically the longer the training is, the better the quality is, but there is also a large gap in image quality within the same iteration. To solve this problem, inspired by previous work [23], we propose employing multiple binary classifiers to learn the GIQA, which we call MBC-GIQA. To be specific, N binary classifiers are trained. For the i-th classifier, the training data is divided into positive or negative samples according to a threshold T^i, given an image $\mathcal{I} \in \mathcal{D}$, its label c^i for the i-th classifier is:

$$c^i = \begin{cases} 0 & \textit{if } S_p(\mathcal{I}) < T^i \\ 1 & \textit{otherwise} \end{cases}, \tag{2}$$

where $i = 1, 2, \ldots, N$ and $0 < T^1 < T^2 < \cdots < T^N = 1$. So a quality score $S_p(\mathcal{I})$ can be converted to a set of binary labels $\{c^1, c^2, \ldots, c^N\}$. Each binary classifier learns to distinguish whether the quality value is larger than T^i. Suppose the predicted score for i-th binary classifier is $\hat{c}^i$, $i = 1, 2, \ldots, N$. So the training loss

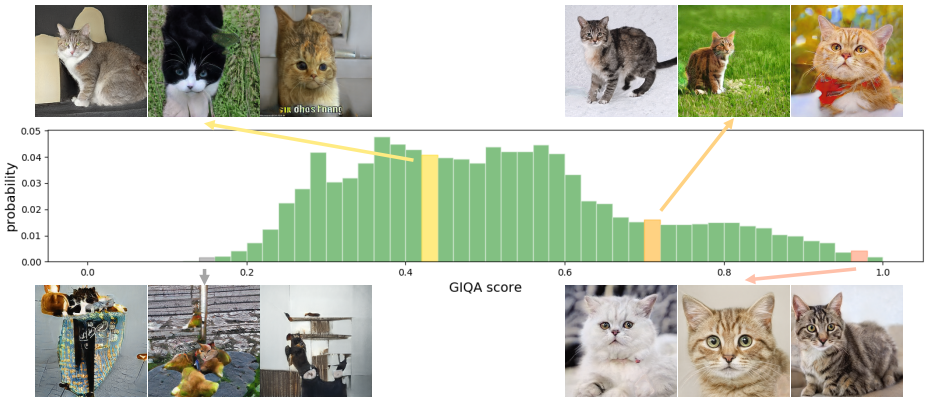

Fig. 2: Generated images from StyleGAN pretrained on LSUN-cat dataset, sorted by GMM-GIQA method. We randomly sample images from different score for better visualization.

for the framework is:

$$L = -\sum_{I \in \mathcal{D}} \sum_{i=1}^{N} (c^i \log(\hat{c}^i) + (1 - c^i) \log(1 - \hat{c}^i)). \tag{3}$$

Using classification instead of regression in this way can be more robust to noise. Although both positive and negative training samples contain noise, T^i is still statistically the decision boundary of i-th classifier. During the inference time, suppose we get all the predicted scores $\hat{c}^i, i = 1, 2, \ldots, N$ for a generated image $\mathcal{I}_g$. Then the final predicted quality score for $\mathcal{I}_g$ is the average of all predicted scores:

$$S_{\text{MBC}}(\mathcal{I}_g) = \frac{1}{N} \sum_{i=1}^{N} \hat{c}^i. \tag{4}$$

3.2 Data-Based Methods

Data-based methods aims to solve the quality estimation in a probability distribution perspective. We directly model the probability distribution of the real data, then we can estimate the quality of a generated image by the estimated probability from the model. We propose adopting two density estimation methods: Gaussian Mixture Model (GMM) and K Nearest Neighbour (KNN).

Gaussian Mixture Model: We propose adopting the Gaussian Mixture Model (GMM) to capture the real data distribution for GIQA. We call this method GMM-GIQA. A Gaussian mixture model is a weighted sum of M component Gaussian densities. Suppose the mean vector and covariance matrix for i-th Gaussian component are $\boldsymbol{\mu}^i$ and $\boldsymbol{\Sigma}^i$, respectively. The probability of an image $\mathcal{I}$

is given by:

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} \mathbf{w}^i g(\mathbf{x}|\boldsymbol{\mu}^i, \boldsymbol{\Sigma}^i), \tag{5}$$

where $\mathbf{x}$ is the extracted feature of $\mathcal{I}$. Suppose the feature extractor function is $f(\cdot)$, so $\mathbf{x} = f(\mathcal{I})$. $\mathbf{w}^i$ is the mixture weights, which satisfies the constraint that $\sum_{i=1}^{M} \mathbf{w}^i = 1$. And $g(\mathbf{x}|\boldsymbol{\mu}^i, \boldsymbol{\Sigma}^i)$ is the component Gaussian densities.

The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices, and mixture weights from all component densities. These parameters are collectively represented by the notation, $\lambda = \{\mathbf{w}^i, \boldsymbol{\mu}^i, \boldsymbol{\Sigma}^i\}$. To estimate these parameters, we adopt the expectation-maximization (EM) algorithm [4] to iteratively update them. Since the probability of a generated image represents its quality score, the quality score of $\mathcal{I}_g$ is given by:

$$S_{\mathrm{GMM}}(\mathcal{I}_g) = p(f(\mathcal{I}_g)|\lambda). \tag{6}$$

K **Nearest Neighbour**: When the real data distribution becomes complicated, it would be difficult to capture the distribution with GMM well. In this situation, we introduce a non-parametric method based on K Nearest Neighbor (KNN). We think the Euclidean distance between generated images and nearby real images in feature space could also represent the probability of generated image, suppose the feature of a generated sample is $\mathbf{x}$. Its k-th nearest real sample's feature is $\mathbf{x}^k$. We can calculate the probability of generated image as:

$$p(\mathbf{x}) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{||\mathbf{x} - \mathbf{x}^k||^2}. \tag{7}$$

Suppose the feature extractor function is also $f(\cdot)$. The quality score of $\mathcal{I}_g$ is given by:

$$S_{\mathrm{KNN}}(\mathcal{I}_g) = p(f(\mathcal{I}_g)). \tag{8}$$

Above all, we introduce three approaches to get three forms of quality score function $S(\mathcal{I}_g)$: $S_{\mathrm{MBC}}(\mathcal{I}_g)$, $S_{\mathrm{GMM}}(\mathcal{I}_g)$, and $S_{\mathrm{KNN}}(\mathcal{I}_g)$. We believe these methods will serve as baselines for further research. In terms of recommendation, we recommend the GMM-based method, since this method outperforms other methods(Table 1) and is highly efficient.

4 Applications

The proposed GIQA framework is simple and general. In this section, we will show how GIQA can be applied in many applications, such as generative model evaluation, improving the performance of GANs.

4.1 Generative Model Evaluation

Generative model evaluation is an important research topic in the vision community. Recently, a lot of quantitative metrics have been developed to assess the performance of a GAN model based on the realism and diversity of generated images, such as Inception Score [31] and FID [13]. However, both of them summarise these two aspects. Our GIQA model can separately assess the realism and diversity of generated images. Specifically, we employ the mean quality score from our methods to indicate the realistic performance of the generative model. Supposing the generative model is G, the generated samples are $\mathcal{I}_g^i, i = 1, 2, \ldots, N_g$. So the quality score of generator G is calculated with the mean quality of N_g generated samples:

$$QS(G) = \frac{1}{N_g} \sum_{i}^{N_g} S(\mathcal{I}_g^i), \tag{9}$$

On the other hand, we can also evaluate the diversity of the generative model G. Note that the diversity represents the relative diversity compared to real data distribution. We exchange the positions of real and generated images in data-based GIQA method. We use generated images to build the model and then evaluate the quality of the real images.

Considering if the generated samples have similar distribution with real samples, then the quality of the real samples is high. Otherwise, if the generated samples have the problem of "mode collapse", which means a low diversity, then the probability of the real samples become low. This shows by exchanging the position, the quality of real samples is consistent with the diversity of generative models. Supposing the real samples are $\mathcal{I}_r^i, i = 1, 2, \ldots, N_r$, the score function built with generative model G is $S'(\cdot)$. The diversity score of the generative model is calculated with mean quality of N_r real samples:

$$DS(G) = \frac{1}{N_r} \sum_{i}^{N_r} S'(\mathcal{I}_r^i), \tag{10}$$

In summary, we have the quality score (QS) and diversity score (DS) to measure the quality and diversity of the generative model separately.

4.2 Improve the Performance of GANs

Another important application of GIQA is to help the generative model achieve better performance. In general, the quality of generated images from a specific iteration of the generator have large variance, we can assess the quality of these generated samples by using our GIQA method, then we force the generator to pay more attention to these samples with low quality. To achieve this, we employ online hard negative mining (OHEM) [32] in the discriminator to apply a higher loss weight to the lower quality samples. To be specific, we set a quality threshold T_q. Samples with quality lower than the threshold T_q will be given a large loss weight $w_l > 1$.

4.3 Image Picker Based on Quality

Another important application of GIQA is to leverage it as an image picker based on quality. For the wide applications of generative models, picking high quality generated images is of great importance and makes these applications more practical. On the other hand, for a generative model to be evaluated, we can take full advantage of the image picker to discard these images with low quality to further improve performance.

5 Experiments

In this section, we first introduce the overall experiment setups and then present extensive experiment results to demonstrate the superiority of our approach.

Datasets and Training Details We conduct experiments on a variety of generative models trained on different datasets. For unconditional generative models, we choose WGAN-GP [12], PGGAN [18], and StyleGAN [19] trained on FFHQ [19], and LSUN [39] datasets. For conditional generative models, we choose pix2pix [15], pix2pixHD [35], SPADE [29] trained on Cityscapes [3] datasets. FFHQ is a large dataset which contains 70000 high-resolution face images. LSUN contains 10 scene categories and 20 object categories, each category contains a large number of images. The Cityscapes dataset is widely used in conditional generative models. In our experiments, we use all the officially released models of these methods for testing.

For learning-based methods, we need the generated images at different iterations of a generative model for training. Specifically, for unconditional generative models, we collect the generated images in the training process of StyleGAN for training, and test the resulting model on the generated images from PGGAN, StyleGAN, and real images. For the conditional generative model, we use the generated images at different iterations of pix2pixHD for training, and test it on the generated images from pix2pix, pix2pixHD, SPADE and real images. To get these training images, we use the official training code, and collect 200,000 generated images, which consist of images from 4000 iterations, 50 images per iteration. We adopt 8 binary classifiers for the MBC-GIQA approach. For the GMM-GIQA method, we set the number of Gaussian components to 7 for LSUN and Cityscapes datasets, and 70 for FFHQ. For the KNN-GIQA method, we set K to 1 for FFHQ and Cityscapes datasets, 3500 for LSUN. All features are extracted from the inception model [33] which is trained on ImageNet. More details please refer to the supplementary material.

Evaluation Metrics Evaluating GIQA algorithms is an open and challenging problem. To quantitatively evaluate the performance of these algorithms, we collect a dataset which is annotated by multiple human observers. To be specific, we first use the generated images from PGGAN, StyleGAN, and real images to build 1500 image pairs[3], then we demonstrate these pairs to 3 human observers

[3] We not only collect images from pretrained models, but also some low quality images from the training procedure.

methods	FFHQ	LSUN-cat	Cityscapes
NIMA[34]	0.598	0.583	0.827
DeepIQA[1]	0.581	0.550	0.763
RankIQA[24]	0.573	0.557	0.780
BC-GIQA	0.663	0.710	0.768
IR-GIQA	0.678	0.784	0.837
SGM-GIQA	0.620	0.829	0.847
MBC-GIQA(**our**)	0.731	0.831	0.886
GMM-GIQA(**our**)	**0.764**	**0.846**	0.895
KNN-GIQA(**our**)	0.761	0.843	**0.898**

Table 1: Comparison of the accuracy on LGIQA dataset for different methods.

to choose which image has a better quality. Finally, we discard the pairs which have inconsistent human opinions. The number of remaining pairs are 974, 1206, and 1102 for FFHQ, LSUN-cat, and Cityscapes dataset, respectively. We refer to this dataset as the Labeled Generated Image Quality Assessment(LGIQA) dataset. To evaluate a GIQA algorithm, we employ the algorithm to rank the image quality in each pairs and check if it is consistent with the annotation. Thus we can calculate the accuracy of each algorithm.

5.1 Comparison with Recent Works

Since no previous approach aims to solve the problem of GIQA, we design several baselines and compare our approach with them to validate our approach.

The first baselines are the methods for natural image quality assessment, we choose recent works like DeepIQA [1], NIMA [34], RankIQA [24] for comparison. For DeepIQA and NIMA, we directly apply their released model for testing. For RankIQA, we use their degradation strategy and follow their setting to train a model on our datasets. The second baselines are related to the learning-based method. We adopt the simple idea of directly employing a CNN network to regress pseudo label of quality score $S_p(\mathcal{I})$, which is called IR-GIQA. Another idea is instead of using multiple binary classifiers, we use only 1 classifier to determine whether the image is real or not, we call this BC-GIQA. The third baseline is to capture the real data probability distribution is to use a single Gaussian model, we call this SGM-GIQA.

We present the results in Table 1. We observe that our proposed GIQA methods perform better than those natural image assessment methods. Meanwhile, the MBC-GIQA gets higher accuracy than the baseline IR-GIQA and BC-GIQA, and GMM-GIQA is also better than the SGM-GIQA model. which demonstrates the effectiveness of our proposed method. Overall, GMM-GIQA achieves the best results, so we use GMM-GIQA for the following experiments.

We qualitatively compare the generated image quality ranking results for our proposed GMM-GIQA and NIMA in Fig 3, we observe that GMM-GIQA achieves a better generated image quality ranking results that is more consistent with human assessment. More results can be found in supplemental material.

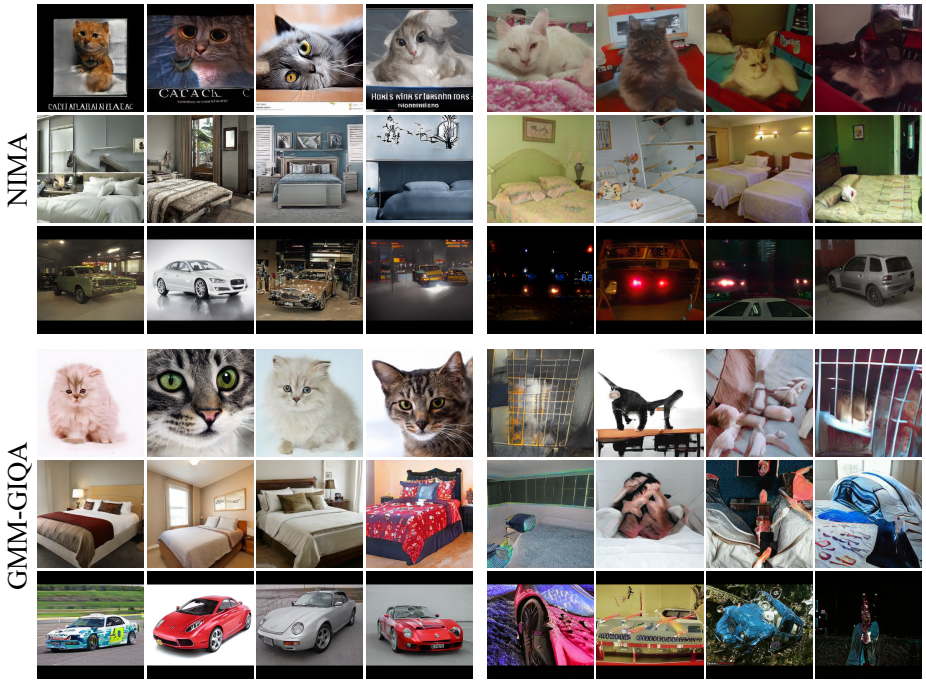

Fig. 3: Generated image quality assessment results for NIMA(the top 3 rows) and our proposed GMM-GIQA(the bottom 3 rows) on LSUN-cat, LSUN-bedroom and LSUN-car datasets. The left are the top-4 high quality images and the right are top-4 low quality images.

5.2 Generative Model Assessment

Quality Distribution Evaluation The proposed GIQA methods are able to assess the quality for every generated sample. Therefore we first employ our proposed GMM-GIQA to validate the quality distribution of generated samples from several generative models. For unconditional generative models, we choose WGAN-GP, PGGAN, StyleGAN trained on FFHQ, LSUN-cat and LSUN-car datasets. For conditional generative models, we choose pix2pix, pix2pixHD, and SPADE trained on Cityscapes dataset. Each generative model generates 5000 test images, and then apply our GMM-GIQA method to calculate the quality score, the quality score distributions are shown in Figure 4. Note that all the quality scores are normalized to $[0, 1]$. We find that PGGAN and StyleGAN are much better than WGAN-GP, and StyleGAN is better than PGGAN. SPADE and pix2pixHD are much better than pix2pix, SPADE is slightly better than pix2pixHD. All these observations are consistent with human observation.

QS and DS for Generative Models As we introduced in Section 4.2, we propose two new metrics the quality score (QS) and diversity score (DS) to

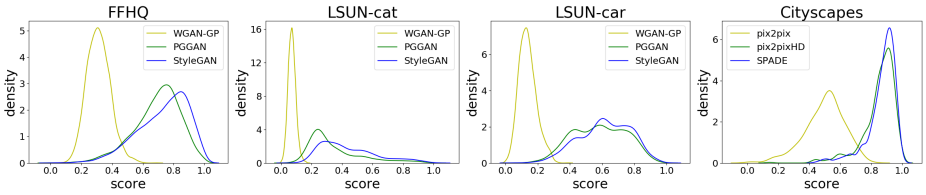

Fig. 4: Quality score distribution of generated images from different generative models.

	FFHQ			LSUN-cat			LSUN-car		
	[12]	[18]	[19]	[12]	[18]	[19]	[12]	[18]	[19]
FID	107.6	14.66	10.54	192.2	49.87	18.67	146.5	14.73	12.70
Prec	0.006	0.640	0.704	0.012	0.487	0.608	0.022	0.608	0.680
Rec	0	0.452	0.555	0	0.356	0.467	0.002	0.487	0.531
QS	0.312	0.694	0.731	0.072	0.347	0.441	0.138	0.583	0.617
DS	0.355	0.815	0.806	0.236	0.789	0.796	0.281	0.801	0.791

Table 2: Comparison of FID, Precision [21], Recall [21], QS, and DS metric for the generative model WGAN-GP [12], PGGAN [18], and StyleGAN [19] on three different datasets: FFHQ, LSUN-cat, and LSUN-car.

assess the performance of generative models. So we compare these two metrics with the FID and [21]. Table 2 reports the results on WGAN-GP, PGGAN, and StyleGAN. Our QS and DS metrics are consistent with other metrics. Also to validate our metric can be applied for conditional models, we report the QS score 0.498, 0.851, and 0.879 for conditional generative models: pix2pix, pix2pixHD and SPADE, respectively. The results are consistent with human observation.

5.3 Improving the Performance of GANs

One important application of GIQA is to improve the performance of GANs. We find that we can achieve this in two ways, one is to adopt the GIQA to discard low quality images from all the generated images for evaluation. The other one is to take full advantage of the GIQA to achieve OHEM in the training process of GANs, then the performance gets improved.

Image Picker Trick We conduct this experiment on the StyleGAN model trained on LSUN-cat. We first generate 10000 images, then we use the GMM-GIQA method to rank the quality of these images and retain different percentages of high quality images, finally we randomly sample 5000 remaining images for evaluation. We test the generated images with different remaining rates. For comparison, we notice that StyleGAN adopt a "truncation trick" on the latent space which also discards low quality images. With a smaller truncation rate, the quality becomes higher, the diversity becomes lower. We report the FID, QS, DS results of different truncation rate and remaining rate in Table 3. We notice that the FID improves when the truncation rate and remaining rate are set to

Methods	Metrics	1	0.9	0.8	0.7	0.6	0.5
truncation rate	FID	18.67	18.05	19.64	23.46	30.48	41.68
	QS	0.441	0.463	0.486	0.510	0.537	0.567
	DS	0.796	0.771	0.756	0.731	0.699	0.686
remaining rate	FID	18.67	16.65	17.63	20.73	25.84	33.19
	QS	0.441	0.466	0.495	0.520	0.551	0.587
	DS	0.796	0.792	0.780	0.766	0.746	0.712

Table 3: Comparison of truncation trick and image picker trick using StyleGAN on LSUN-cat dataset.

Datasets	Methods	FID	QS	DS
FFHQ	StyleGAN	17.35	0.697	0.753
	StyleGAN+OHEM	16.89	0.711	0.755
	StyleGAN+OHEM+Picker	16.68	0.723	0.749
LSUN-cat	StyleGAN	18.67	0.441	0.796
	StyleGAN+OHEM	18.12	0.462	0.790
	StyleGAN+OHEM+Picker	16.25	0.482	0.785

Table 4: Performance comparison of various settings for StyleGAN.

0.9, and the remaining rate works better than the truncation rate. Which also perfectly validates the superiority of the QS and DS metric.

OHEM for GANs To validate whether OHEM improves the performance of GANs, we train two different settings of StyleGAN on FFHQ and LSUN-cat datasets at 256×256 resolution. One follows the original training setting (denoted as StyleGAN), the other applies the OHEM in the training process and puts a large loss weight w_l on low quality images whose quality score is lower than threshold T_q, which we called StyleGAN+OHEM. We set the w_l, T_q to 2, 0.2 in our experiments. After finishing the training process, we evaluate the FID, QS, and DS metric. Table 4 reports the results. We find that OHEM improves the performance of GANs. Besides, based on this model, by using our image picker trick (denoted as StyleGAN+OHEM+Picker), it can further achieve better performance.

5.4 Analysis of the Proposed Methods

In this subsection, we conduct experiments to investigate the sensitiveness of hyper parameters in the proposed three approaches. All the results are evaluated on our LGIQA dataset.

Hyper parameters for MBC-GIQA For MBC-GIQA, what we mainly want to explore is how the number of binary classifiers influence the results, we train the model using different numbers of binary classifiers. The number of classifiers is set to 1, 4, 6, 8, 10, 12. Table 5 reports the results. We find that as the number of binary classifiers increases from 1 to 8, the performance becomes better and better, and as the number continues to increase to 12 the performance degrades.

N	1	4	6	8	10	12
Accuracy	0.663	0.682	0.722	**0.731**	0.718	0.717

Table 5: Results of different number of binary classifiers N for MBC-GIQA.

M	5	10	20	30	50	70	100
Accuracy	0.648	0.663	0.738	0.733	0.752	**0.764**	0.753

Table 6: Results of different number of Gaussian components M for GMM-GIQA.

K	1	30	100	500	1000	2000	3500	5000	7000
Accuracy	0.823	0.828	0.833	0.837	0.840	0.842	**0.843**	0.841	0.840

Table 7: Results of different number of nearest neighbors K for KNN-GIQA.

Hyper parameters of GMM-GIQA The key factor for GMM is the number of Gaussian components M, therefore we explore how M affects the results of GIQA. We set M to 5, 10, 20, 30, 50, 70, 100 and test the results on our LGIQA-FFHQ dataset. We show the results in Table 6, as the number of Gaussian components M increases from 5 to 70, we get better and better results, and the number continues to increase to 100, the performance degrades.

Hyper parameters of KNN-GIQA To explore how the number of nearest neighbours K affects the results, we apply different K in the KNN-GIQA. Specifically, we set K to 1, 30, 100, 500, 1000, 2000, 3500, 5000, 7000. The results on LGIQA-LSUN-cat dataset are shown in Table 7, as K increases from 1 to 1000, we get better and better results, and the number continues to increase to 7000, the performance is comparable.

6 Conclusions

In this paper, we aim to solve the problem of quality evaluation of a single generated image and propose the new research topic: GIQA. To tackle this problem, we propose three novel approaches from two perspectives: learning-based and data-based. Extensive experiments show that our proposed methods can perform quite well on this new topic, also we demonstrate that GIQA can be applied in a wide range of applications.

We are also aware that there exist some limitations of our methods. For the learning-based method MBC-GIQA, it requires the generated images at different iterations for training, while these images may not be easily obtained in some situations. For the data-based method GMM-GIQA, there is a chance of failure when the real data distribution is too complicated. We also notice that our current results are far from solving this problem completely. We hope our approach will serve as a solid baseline and help support future research in GIQA.

References

1. Bosse, S., Maniry, D., Wiegand, T., Samek, W.: A deep neural network for image quality assessment. In: 2016 IEEE International Conference on Image Processing (ICIP). pp. 3773–3777. IEEE (2016)
2. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)
3. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
4. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society: Series B (Methodological) **39**(1), 1–22 (1977)
5. Doersch, C.: Tutorial on variational autoencoders. arXiv preprint arXiv:1606.05908 (2016)
6. Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., Greenspan, H.: Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. Neurocomputing **321**, 321–331 (2018)
7. Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., Greenspan, H.: Synthetic data augmentation using gan for improved liver lesion classification. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). pp. 289–293. IEEE (2018)
8. Golestaneh, S.A., Chandler, D.M.: No-reference quality assessment of jpeg images via a quality relevance map. IEEE Signal Processing Letters **21**(2), 155–158 (2013)
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
10. Gu, S., Bao, J., Yang, H., Chen, D., Wen, F., Yuan, L.: Mask-guided portrait editing with conditional gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3436–3445 (2019)
11. Gu, S., Chen, C., Liao, J., Yuan, L.: Arbitrary style transfer with deep feature reshuffle. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8222–8231 (2018)
12. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: Advances in neural information processing systems. pp. 5767–5777 (2017)
13. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in neural information processing systems. pp. 6626–6637 (2017)
14. Huynh-Thu, Q., Ghanbari, M.: Scope of validity of psnr in image/video quality assessment. Electronics letters **44**(13), 800–801 (2008)
15. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
16. Kang, L., Ye, P., Li, Y., Doermann, D.: Convolutional neural networks for no-reference image quality assessment. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1733–1740 (2014)
17. Kang, L., Ye, P., Li, Y., Doermann, D.: Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks. In: 2015 IEEE international conference on image processing (ICIP). pp. 2791–2795. IEEE (2015)

18. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
19. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2019)
20. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. arXiv preprint arXiv:1912.04958 (2019)
21. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T.: Improved precision and recall metric for assessing generative models. In: Advances in Neural Information Processing Systems. pp. 3927–3936 (2019)
22. Lin, K.Y., Wang, G.: Hallucinated-iqa: No-reference image quality assessment via adversarial learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 732–741 (2018)
23. Liu, T.J., Liu, K.H., Liu, H.H., Pei, S.C.: Age estimation via fusion of multiple binary age grouping systems. In: 2016 IEEE International Conference on Image Processing (ICIP). pp. 609–613. IEEE (2016)
24. Liu, X., van de Weijer, J., Bagdanov, A.D.: Rankiqa: Learning from rankings for no-reference image quality assessment. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1040–1049 (2017)
25. Moorthy, A.K., Bovik, A.C.: A two-step framework for constructing blind image quality indices. IEEE Signal processing letters **17**(5), 513–516 (2010)
26. Odena, A., Dumoulin, V., Olah, C.: Deconvolution and checkerboard artifacts. Distill (2016). https://doi.org/10.23915/distill.00003, http://distill.pub/2016/deconv-checkerboard
27. Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., et al.: Conditional image generation with pixelcnn decoders. In: Advances in neural information processing systems. pp. 4790–4798 (2016)
28. Pan, D., Shi, P., Hou, M., Ying, Z., Fu, S., Zhang, Y.: Blind predicting similar quality map for image quality assessment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6373–6382 (2018)
29. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2337–2346 (2019)
30. Saad, M.A., Bovik, A.C., Charrier, C.: Blind image quality assessment: A natural scene statistics approach in the dct domain. IEEE transactions on Image Processing **21**(8), 3339–3352 (2012)
31. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: Advances in neural information processing systems. pp. 2234–2242 (2016)
32. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 761–769 (2016)
33. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2015)
34. Talebi, H., Milanfar, P.: Nima: Neural image assessment. IEEE Transactions on Image Processing **27**(8), 3998–4011 (2018)
35. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8798–8807 (2018)

36. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
37. Wang, Z., Simoncelli, E.P.: Reduced-reference image quality assessment using a wavelet-domain natural image statistic model. In: Human Vision and Electronic Imaging X. vol. 5666, pp. 149–159. International Society for Optics and Photonics (2005)
38. Yan, Q., Xu, Y., Yang, X.: No-reference image blur assessment based on gradient profile sharpness. In: 2013 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB). pp. 1–4. IEEE (2013)
39. Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., Xiao, J.: Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365 (2015)
40. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. arXiv preprint arXiv:1805.08318 (2018)