A Lightweight Generalizable Evaluation and Enhancement Framework for Generative Models and Generated Samples

Ganning Zhao [1], Vasileios Magoulianitis [1], Suya You [2], C.-C. Jay Kuo [1]
University of Southern California, Los Angeles, California, USA[1]
DEVCOM Army Research Laboratory, Los Angeles, California, USA[2]

Abstract

While extensive research has been conducted on evaluating generative models, little research has been conducted on the quality assessment and enhancement of individual-generated samples. We propose a lightweight generalizable evaluation framework, designed to evaluate and enhance the generative models and generated samples. Our framework trains a classifier-based dataset-specific model, enabling its application to unseen generative models and extending its compatibility with both deep learning and efficient machine learning-based methods. We propose three novel evaluation metrics aiming at capturing distribution correlation, quality, and diversity of generated samples. These metrics collectively offer a more thorough performance evaluation of generative models compared to the Fréchet Inception Distance (FID). Our approach assigns individual quality scores to each generated sample for sample-level evaluation. This enables better sample mining and thereby improves the performance of generative models by filtering out lower-quality generations. Extensive experiments across various datasets and generative models demonstrate the effectiveness and efficiency of the proposed method.

1. Introduction

Generative AI has rapid advancements and finds extensive applications across a range of domains, including image generation, image inpainting, and image-to-image translation. Given the burgeoning prevalence of generative models, it is imperative to devise solutions for an automated and objective quality assessment of generated samples.

A large number of quantitative metrics have been proposed in recent years [4, 5], with notable examples encompassing the Inception Score (IS) [33], Fréchet Inception Distance (FID) [11], Classifier two-sample test [27, 54], and Precision and Recall (P&R) [32]. Each of these metrics has its own strengths and limitations. However, these popular evaluation methodologies share three main issues. Firstly, they evaluate generative models based on the statistical aggregation of generated samples, precluding their applicability for evaluating individual samples and enhancing the generative models. Secondly, the computational complexity of these methodologies is huge, since they rely on deep features from late layers of deep neural networks (DNNs). Moreover, a number of these evaluation methodologies exhibit a bias towards the object dataset, ImageNet, a dataset predominantly used for pre-training networks. Despite ongoing endeavors to devise enhanced quality evaluation methodologies [1,12,30], these fundamental challenges persist.

Recently proposed classifier-based methods [27, 51, 52, 54] present a set of evaluation metrics for generative models and generated samples. However, these methods demand more computational resources to train the evaluation framework per generative model. Therefore, these methods have limited generalizability across various generative models within the datasets. Additionally, these metrics provide evaluations at the distribution level and lack the capability to assess the diversity of the generated samples. Furthermore, little methodology evaluates the quality of individual generated samples, and the role of evaluation models in augmenting the performance of generative models is unexplored.

To address these challenges, we propose an evaluation framework to enhance the generalizability to unseen generative models and reduce the computational demands in practical applications. This novel framework is lightweight and applicable to sample-level evaluation, and more versatile, extending its applicability to diverse tasks within the realm of generative models, including image generation, image completion [54], and image-to-image translation [36,50,53]. In particular, we incorporate a dataset construction module into the classifier-based framework, empowering it to learn from a variety of generative models and subsequently apply this knowledge to unseen generative models. Additionally, we integrate both lightweight machine learning and efficient deep learning-based feature

extractors, further enhancing the framework's generalizability and efficiency. Additionally, we introduce both the evaluation and enhancement of generative models within the framework, ensuring a comprehensive and systematic presentation of these crucial aspects. Furthermore, we propose three evaluation metrics to comprehensively assess the generative model performance, including accuracy, diversity, and quality.

The rest of this paper is systematically organized as follows: Section 2 provides an overview of related work, Section 3 details the methodology, Section 4 showcases experimental results, and Section 5 concludes the paper.

2. Related Work

Existing literature has been prolific during the past decade about evaluation metrics of generative models. Most of the proposed metrics aim at measuring the distance of the distributions between the generated and the real samples.

In Section 2.1, we review the most commonly used evaluation metrics, as well as some works from the sample-based quality assessment. All of those works are built upon the learned feature space of Deep Neural Networks (DNNs). Since our work employs the Efficient Machine Learning framework for creating a feature space embedding for distance measuring, we review its basic elements in Section 2.2.

2.1. AI Generated Image Quality Metrics

Inception Score (IS) was proposed by Salimans *et al.* [33] as a better optimization technique for training GANs. One significant limitation of IS is that it lacks capturing the intra-class diversity of the generated samples. Gurumurthy *et al.* [10] proposed a modified version of IS, trying to quantify the content diversity, even with limited training data, by using mixture models and their joint optimization along with the GAN parameters.

Another commonly used metric is the Fréchet Inception Distance (FID) proposed by Heusel *et al.* [11]. It is based on measuring the distance between synthetic and generated samples under the assumption they are normally distributed. It employs the Inception-Net-V3 for creating the feature embedding space and calculating its distribution. This approach improves the diversity of generated samples compared to IS. Another work was proposed [25] to further improve the intra-class variability on the FID metric.

On a slightly different approach, Liu *et al.* [28] proposed the precision-recall metric, aiming at quantifying both ends of the problem. Precision captures the distance between the real and generated samples, while recall represents the class diversity. However, the fact that the reference distribution is not usually available, makes this metric impractical [30]. Classifier-based works [12, 27] employ a classifier to work as a discriminator model between fake and real samples, where the classifier's accuracy is used as the evaluation metric.

All the previously mentioned works ultimately aim at evaluating the performance of generative models and help in their optimization. They lack the ability to evaluate the quality of individually generated samples. This can lead to a better optimization for GANs, by enabling online hard negative mining or in other applications where we need to assess the quality of synthetic samples, for rejecting those of poor quality. An earlier work [3] has provided definitions of image distortion and generated sample quality, as well as realized the tradeoff between distortion and perceptual quality.

The Generated Image Quality (GIQA) term is introduced by Qu *et al.* [9] aiming at evaluating individual samples and realizing a quality index, that can correlate well with human perception with respect to the quality and content of the sample. On the same line of research, another work [38] was earlier proposed, also deploying a DNN for predicting a quality score with no reference. In the context of large-scale image completion, a new metric named Paired/Unpaired Inception Discriminative Score (P-IDS/U-IDS) was proposed by Zhao *et al.* [54]. It measures the perceptual quality of inpainted images which correlates with human opinion, by using a DNN for feature extraction and a Support Vector Machine (SVM) with a linear kernel to fit real and fake samples out of the deep feature space. According to their results, P-IDS achieves a high correlation coefficient with human preference rate. Recently, a dataset [24] for generated image quality assessment has been released, namely AGIQA-3K, which comprises fine-grained subjective scores for generated images from six different models.

2.2. Efficient Machine Learning

Decoupling from other approaches that use DNNs for mapping the input image samples onto an embedding feature space, this work uses the Efficient Machine Learning (Green Learning, GL) paradigm [20] for feature extraction. The framework consists of a multi-scale feed-forward linear model that learns features in an unsupervised way with no back-propagation. Hence, features are independent from the objects and classes of the training set, making them a bias-free approach from the pre-trained dataset's content. Besides, one of the key benefits behind GL is the low complexity and model size, making easier its deployment across several devices. Also, the transparent feature extraction process is another advantage for more interpretability within the feature extraction process.

In the green learning framework, there are a couple of models for feature extraction. E-Pixelhop [44] is a popular choice for different image classification tasks. It creates a rich spatial-spectral representation of the input image at different scales, by using subspace approximation that

maximizes the data variance. Within the GL framework a feature selection method, namely Discriminant Feature Test (DFT) [45] has been proposed to filter out irrelevant features for the targeted task. The remainder most discriminant feature subset can be used for measuring the quality of generated samples, in terms of its discriminant spatial-spectral components.

Although GL is an emerging framework, it has been already successfully applied in several problems, such as image quality assessment [29, 49], point cloud classification, segmentation and registration [13–15, 26, 47], image enhancement [2], texture synthesis [22, 23], as well as graph node classification [42, 43]. Our method is built upon the learned GL feature space for the generated sample quality evaluation.

3. Proposed Method

In this paper, we introduce a comprehensive framework designed to assess either the performance of generative models or the quality of the samples they generate, ultimately contributing to the improvement of the generative models. The architecture of our proposed method is depicted in Fig. 1, and it comprises four distinct components, detailed in the subsequent sections.

3.1. Dataset Construction

The core concept of dataset construction revolves around developing a dataset-specific evaluation model, universally applicable to all generative models associated with that particular dataset. Given a dataset and k distinct generative models $GM_1, GM_2, ..., GM_k$, and N groundtruth (real) images, we construct the training set using samples generated by the randomly selected first i generative models $GM_1, GM_2, ..., GM_i$, where $i < k$. We label all generated samples as "0" and real samples as "1". To eliminate class imbalance, we randomly select N/i samples from each generative model. Therefore, the number of training samples including ground truth and generated samples is $2N$. We construct a test set using samples generated by the remaining generative models $GM_{i+1}, GM_{i+2}, ..., GM_k$, as well as unseen samples produced by $GM_1, GM_2, ..., GM_i$. The evaluation of the generative model under investigation is conducted by feeding its generated samples in the test set.

3.2. Feature Extractor

To validate the generalizability of our framework across both deep learning and lightweight machine learning methodologies, we introduce two distinct options for feature extraction, each with its unique advantages and limitations. We provide a detailed discussion of each method below:

3.2.1 Deep Learning-Based Feature Extraction

Our framework is compatible with a variety of deep learning networks for feature extraction, such as Inception-V3, VGG, ResNet, and EfficientNet, etc. Deep features have demonstrated a strong correlation with human perception [48], and show high performance. On the other hand, the substantial number of parameters and FLOPs of networks require significant computational resources. Additionally, since all these models are pre-trained on the ImageNet dataset, their features are inherently biased towards object datasets, potentially leading to suboptimal performance when applied to other datasets, such as the LSUN-bedroom dataset, LSUN-church dataset, and grayscale images.

3.2.2 Machine Learning-Based Feature Extraction

To overcome those challenges while maintaining performance by learning visual-correlated and discriminative features. This module consists of two submodules, as elaborated below:

1. Representation Learning

We aim to jointly learn effective local spatial and global spectral representations of images. We analyze overlapping blocks of the images, by applying the Saab transform [21] to learn meaningful local representations. This transform helps in extracting both low and high-frequency components from the image blocks, converting them into 3D tensors of spatial representation. To reduce dimensionality and computational load, we discard high-frequency components with negligible energy. Absolute max-pooling is then applied to enhance the robustness of the spatial representation, which also serves as the input for the next global processing stage.

Following this, we apply the channel-wise Saab transform [7] to further reduce dimensions and capture longer-distance correlations effectively. This step generates spectral representations of the image. Finally, we concatenate these spatial local and spectral global representations, creating a rich feature set for discriminant feature selection in the subsequent module. This approach ensures a balance between detail capture and computational efficiency, providing a robust foundation for representation analysis.

2. Feature Selection

To derive the most relevant set of features learned in the representation module, we employ the Discriminant Feature Test (DFT) [45], a method designed to quantify and rank the discriminative power of each feature. The DFT method partitions the value range of each feature dimension, calculating the DFT loss to measure its discriminative capability. Lower DFT loss indicates higher relevance for the task at hand and thus has higher discriminative power.

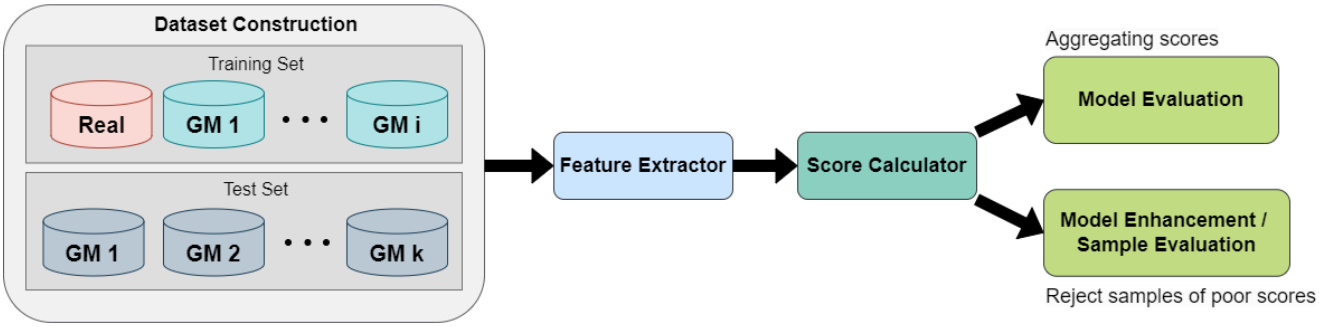


Figure 1. The overview of the proposed method. The framework comprises four components, including dataset construction, feature extraction, score calculation, and task-specific operation. Each cylinder in the dataset construction module represents the generated samples by certain generative models (GM). In the task-specific module, different operations are conducted for model evaluation, model enhancement, or sample evaluation.

Leveraging the independence of representations, we compute the DFT loss in each dimension, subsequently ranking the features based on this metric. An elbow point in the DFT loss curve then guides the selection of the most discriminative subset of features, which are forwarded to a binary classifier in Sec. 3.3 for subsequent analysis.

Manipulating Higher-Resolution Images

To avoid significant information loss during image downsampling for higher-resolution datasets, we introduce a multi-scale pipeline that comprises two branches with identical representation learning and feature selection architectures for global and local feature learning. Images are first downsampled to a reasonable size. Local crops extracted from images are fed into the local learning branch, while global crops are further downsampled as the input to the global learning branch. This setup enables simultaneous feature extraction of low-resolution global layouts (global branch) and high-resolution local details (local branch). Features learned from these two branches will subsequently be forwarded into the following module for further analysis and processing.

3.3. Score Calculator

We evaluate the quality of each generated sample, assigning it a quantitative score as its quality index. Compared with FID, binary classification captures subtle variations, exhibits a stronger correlation with human perception, and requires fewer training samples to converge [27, 54]. Furthermore, the binary classifier can be applied to assess the quality of individual samples.

Therefore, we employ a binary classifier within this module, training it to distinguish between real and generated samples. In the training stage, generated samples are labeled as "0", while real samples are "1". During the inference phase, the classifier assigns a soft label, $0 \leq \hat{p} \leq 1$, to each sample. A soft label closer to 0.5 indicates a more difficult differentiation between real and generated samples, so it closely resembles real samples in the feature space and, consequently, is of higher quality. Similarly, a soft label near the tails, "0" or "1" indicates lower quality. Therefore, we utilize the soft label as a proxy for the quality score of each generated sample, providing a reliable metric for evaluation.

3.4. Model Evaluation and Enhancement

This section depicts two pathways for the enhancement and evaluation of generative models, dependent upon the specific requirements of the task.

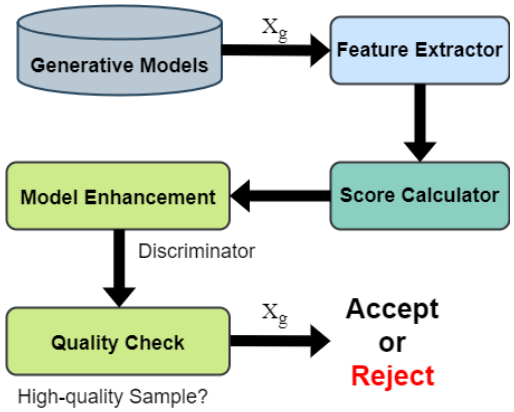


Figure 2. Model enhancement. We get the evaluation scores by feeding the generated samples to our framework. Generative models are enhanced by filtering out samples of evaluated low quality.

Model Enhancement

This procedure involves the exclusion of poor generations assessed by our scoring calculator. Specifically, we sort the scores assigned to all generated samples under test in ascending order. Samples of higher quality are typically

located in regions of the feature space where distinguishing between real and generated samples becomes challenging for a classifier, resulting in a soft label close to 0.5, or the chance level. Consequently, we discard samples at the tails of this distribution, which are of inferior quality, resulting in a refined collection of higher-quality samples.

Model Evaluation

We aggregate the scores across all generated samples and proceed to calculate three performance metrics: accuracy, quality, and diversity. Accuracy (Acc.) measures the correlation between the distributions of real and generated samples in the feature space, quantified as the ratio of correct decisions to the total number of decisions made. A lower accuracy score implies better generative models, indicating a closer distribution correlation and difficulty in distinguishing between the distributions of real and generated samples, approaching a chance level, close to 0.5.

Quality and diversity are metrics explained using Figure. 3. Mathematically, we consider the generated class as positive (P) and the real class as negative (N). Quality is calculated as the probability of positive samples incorrectly predicted as negatives, divided by all positive samples:

$$f_{quality} = \frac{FN}{FN+TP} = 1 - \frac{TP}{FN+TP}$$

where TP, FP, TN, and FN denote true positive, false positive, true negative, and false negative, respectively. Thus, the quality index can be computed as $f_{quality} = 1 - Recall_g$ where $Recall_g$ indicates the recall of the generated class.

We consider the real class as positive (P) and the generated class as negative (N), diversity is computed as the probability of negative samples incorrectly predicted as positive, divided by all samples predicted as positive:

$$f_{diversity} = \frac{FP}{FP+TP} = 1 - \frac{TP}{FP+TP}$$

Therefore, the diversity index is indicated by $f_{diversity} = 1 - Precison_r$ where $Precison_r$ denotes the precision of the real class. Higher quality and diversity scores are preferable.

4. Experiments

In this section, we present comprehensive experiments of our proposed method across various datasets and generative models, demonstrating its effectiveness and versatility.

4.1. Experimental Setup

Datasets

Our experiments span across four diverse datasets to validate the robustness and generalizability of our method: CIFAR-10 [19] (32x32 resolution images), LSUN-Church, LSUN-Bedroom [46] (both with 256x256 resolution images), and LSUN-10 [8] (96x96 resolution images).

Generative Models

We evaluate the quality of generated samples and compare and enhance the performance across multiple generative models, including Diffusion-StyleGAN2 [41], Diffusion-ProjectedGAN [41], StyleGAN2-ADA [17], StyleFormer [31], StyleGANXL [35], E2GAN [40], StyleGAN [18], ProgressiveGAN [16], and ProjectedGAN [34].

Implementation details

Data Construction. We split the generative models into training and test sets, ensuring test sets comprise untouched generative models.

Feature Extractor. For deep learning-based feature extraction, we use EfficientNet [39], specifically the EfficientNet-B1. This choice is motivated by its small number of parameters and FLOPs, aligning with our lightweight objective. Additionally, comparative studies indicate that EfficientNet-B1 exhibits performance slightly superior to the widely-used Inception-V3 [37], a standard feature extractor in performance evaluation metrics such as FID [11] and IS [33], while it is more lightweight. For machine learning-based feature extraction, we adhere to the parameter settings outlined in [52].

Score Calculator. We employ the XGBoost (extreme gradient boosting) classifier [6] due to its proven high performance in various applications, especially within the green learning framework.

4.2. Results and Analysis

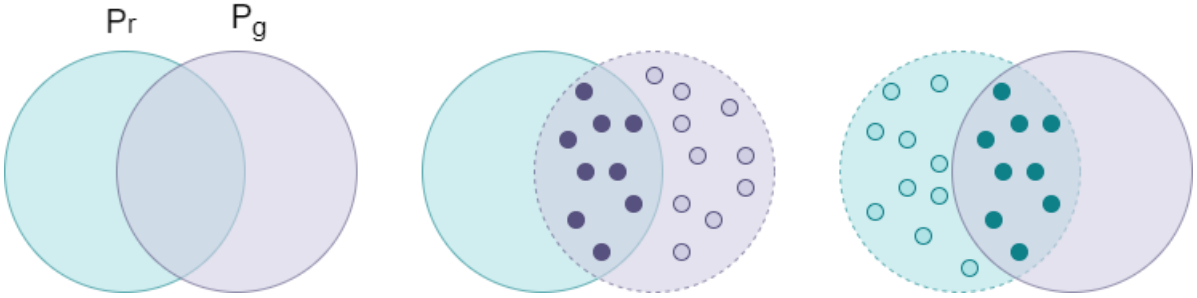
4.3. Sample Evaluation

Figure 4 displays the histogram of quality scores (soft labels) for samples generated by Styleformer with a machine learning-based feature extractor. Compared to Diffusion-StyleGAN2, Styleformer exhibits superior performance on the CIFAR-10 dataset but inferior on the STL-10 dataset. It leads to a higher concentration of samples with probabilities near 0.5 in (a) as real and generated samples are harder to differentiate.

Figure 5 illustrates the correlation between the quality assessments from our framework and human visual perception. Samples evaluated as high-quality by our framework, exhibiting soft scores near 0.5, indeed appear visually superior compared to those evaluated as low-quality, which have soft scores close to ”0” or ”1”.

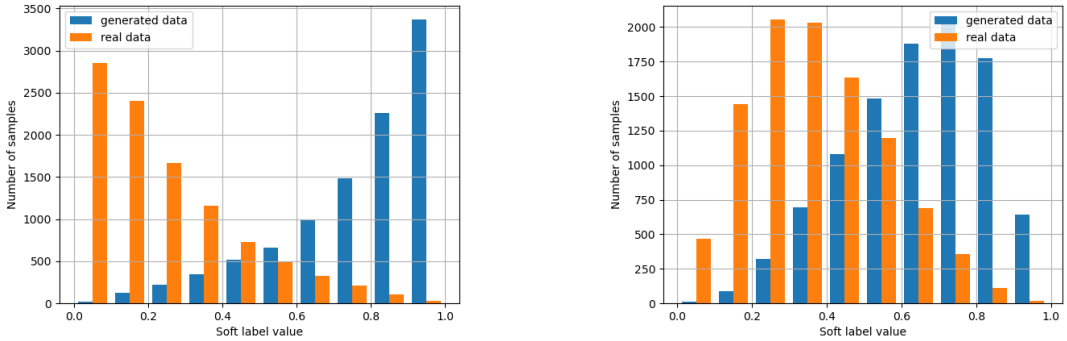
4.4. Model Evaluation

Beyond serving as a measure of quality on a per-sample basis, our framework also offers evaluation metrics for generative models through the aggregation of quality indices from the generated samples. As elaborated in Section 3.4,



(a) Distribution (b) Quality (c) Diversity

Figure 3. (a) indicates the distribution of real images, denoted as P_r, and the distribution of generated images, represented as P_g. (b) Quality refers to the likelihood that an arbitrary image from P_g lies within the range of P_r. (c) Diversity measures the chance that a random image from P_r is encompassed by the range of P_g.



(a) (b)

Figure 4. The soft label histograms of samples generated by Styleformer and real samples on (a) CIFAR-10 and (b) STL-10 datasets.

classification accuracy (Acc.), quality, and diversity are the three metrics utilized for assessing generative models. We draw comparisons between rankings of the most popular FID metric and those derived from the three metrics on deep learning (EfficientNet-B1) and machine learning-based feature extractors, across various datasets including LSUN-Church, LSUN-Bedroom CIFAR-10, and STL-10, with the results presented in Tables 1, and 2.

The models were initially pre-trained using Progressive-GAN and Diffusion-ProjectedGAN, followed by testing the generative models across four different scenarios: (A), (B), (C), and (D), on the LSUN-Bedroom and LSUN-Church datasets, respectively. The performance ranking of accuracy and FID is consistent since both of them are distribution-level evaluations. However, the scenario changes when assessing quality and diversity; here, the performance difference among generative models diverges, revealing additional evaluative insights into the diversity of the models that the FID score alone may not capture.

Comparatively, scenarios (B) and (D) outperform (A) and (C) in terms of scores, due to the evaluation models being pre-trained on the generative models of (B) and (D). Nonetheless, even within (B) and (D), the performance ranking between accuracy and FID remains consistent, underscoring the reliability of these metrics across different experimental settings.

Since the image sizes of CIFAR-10 and STL-10 are too small for EfficientNet-B1, we experiment with the machine learning-based method, where we can observe similar evaluation results with LSUN datasets. For each dataset, we cross-evaluate the performance of models (E) and (F): the evaluation results for (E) are obtained using a model pre-trained on the generative models of (F), and vice versa, the results for (F) are derived from a model pre-trained on (E).

4.5. Model Enhancement

The framework can be a discriminator to enhance the performance of generative models by filtering out poorly generated samples. Specifically, we filter out generated samples at the tails of the soft label distribution, targeting those with exceedingly high or low soft labels. As a result, the kept samples are more correlated with real samples,

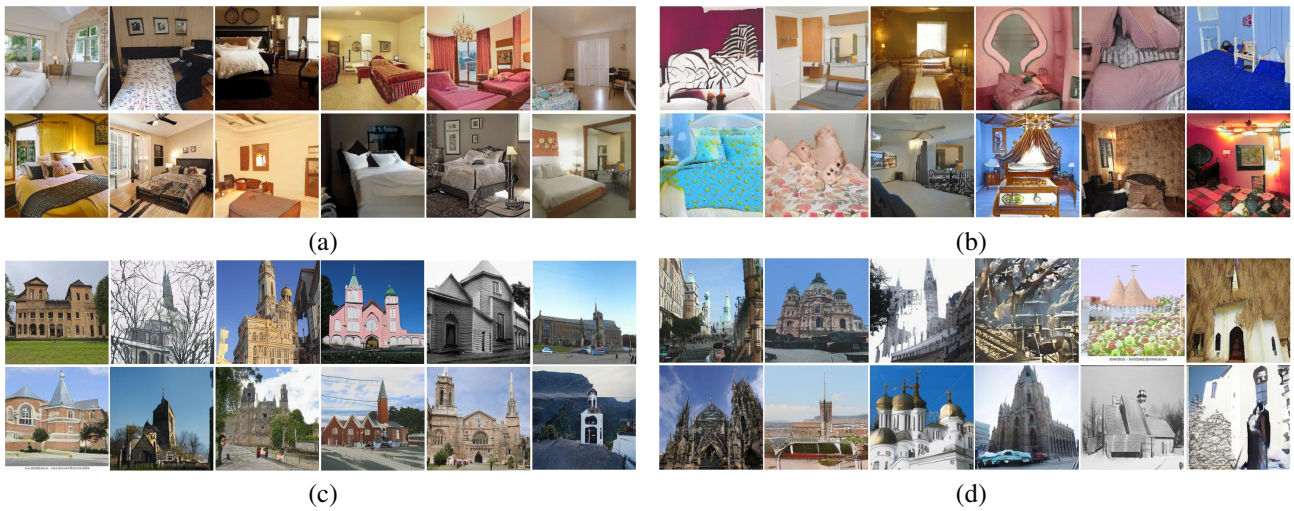


Figure 5. ProjectedGAN-generated samples evaluated as (a) high quality, (b) low quality on LSUN-bedroom dataset and (c) high quality, (d) low quality on LSUN-church dataset by deep learning based method. Apparently, evaluation results are consistent with visual quality.

Table 1. Comparison of four evaluation metrics (FID, Accuracy, Quality, and Diversity) on multiple generative models for LSUN-Bedroom and LSUN-Church datasets. The evaluation model is pre-trained on generative models of (B) and (D) in LSUN-bedroom and LSUN-church datasets, respectively

Settings	Generative Models	FID ↓	EfficientNet-B1			Machine Learning		
			Accuracy ↓	Quality ↑	Diversity ↑	Accuracy ↓	Quality ↑	Diversity ↑
LSUN-Bedroom dataset								
(A)	Diffusion-StyleGAN2	3.65	0.613	0.500	0.408	0.630	0.435	0.307
	StyleGAN	2.65	0.611	0.504	0.410	0.600	0.494	0.416
	ProjectedGAN	1.52	0.583	0.558	0.435	0.593	0.508	0.423
(B)	ProgressiveGAN	8.34	0.793	0.138	0.160	0.717	0.261	0.274
	Diffusion-ProjectedGAN	1.43	0.630	0.465	0.391	0.666	0.363	0.344
LSUN-Church dataset								
(C)	Diffusion-StyleGAN2	3.17	0.610	0.486	0.408	0.778	0.288	0.255
	ProjectedGAN	1.59	0.553	0.600	0.460	0.719	0.407	0.326
(D)	ProgressiveGAN	6.42	0.773	0.159	0.184	0.863	0.120	0.125
	Diffusion-ProjectedGAN	1.85	0.652	0.400	0.362	0.821	0.203	0.194

leading to enhanced performance.

Figure 6 demonstrates the performance improvements achieved by filtering out the poorest generations at varying sampling ratios across different frameworks and datasets. Specifically, we keep fewer generated samples and filter out more bad generations in different ratios. For performance evaluation, we introduce an equal number of real samples to match the quantity of the remaining generated samples. Following this model enhancement, we observe a decrease in accuracy, and an increase in quality and diversity, which indicates improved model performance. Notably, the performance enhancement is more rapid when utilizing deep learning-based models compared to machine learning-based methods, highlighting the efficacy of the former in this context.

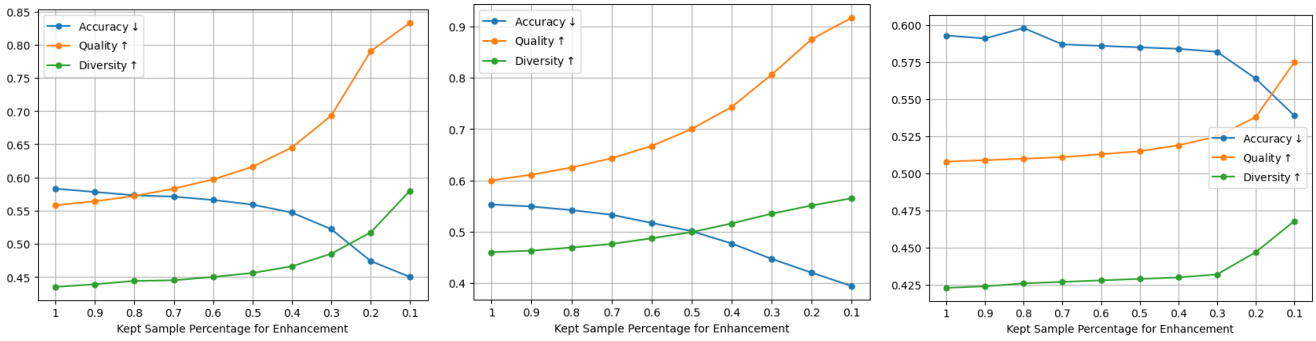
4.6. Efficiency Analysis

We assess the efficiency of machine learning-based feature extractors by comparing their model sizes (in terms of the number of parameters) and computational complexity (measured in floating-point operations, or FLOPs) against popular deep learning-based feature extractors. The majority of leading-edge evaluation techniques extract features using pre-trained Inception-v3, VGG-16, or ResNet-34.

We also evaluate the efficiency in comparison to EfficientNet-B1, which we employ due to its superior accuracy relative to Inception-Net-V3, while maintaining a significantly smaller model size and fewer FLOPs, aligning with our objective for a lightweight solution. The results highlight a substantial disparity in both computational complexity and model size. It is evident that our efficient

Table 2. Comparison of four evaluation metrics (FID, Accuracy, Quality, and Diversity) on multiple generative models for CIFAR-10 and STL-10 datasets. Results in (E) are tested on the model pre-trained on (F), while results in (F) are tested on the model pre-trained on (E).

Generative Models	FID ↓	(E)			Generative Models	FID ↓	(F)		
		Accuracy ↓	Quality ↑	Diversity ↑			Accuracy ↓	Quality ↑	Diversity ↑
CIFAR-10 dataset									
StyleGAN2-ADA	2.92	0.584	0.631	0.442	Diffusion-StyleGAN2	3.19	0.599	0.516	0.420
StyleGAN-XL	1.85	0.516	0.766	0.490	Styleformer	2.82	0.565	0.583	0.450
STL-10 dataset									
E2GAN	25.4	0.881	0.147	0.161	Styleformer	15.2	0.812	0.251	0.253
Diffusion-ProjectedGAN	6.91	0.648	0.576	0.429	Diffusion-StyleGAN2	11.6	0.621	0.601	0.448



(a) (b) (c)

Figure 6. Model enhancement of ProjectedGAN on (a) bedroom dataset, (b) church datasets using deep learning-based feature extractor, and (c) bedroom dataset using our efficient machine learning-based feature extractor. The performance improves when filtering out more bad generations.

Table 3. Comparison of computational complexity and model size of our efficient machine learning and deep learning networks used for feature extractor.

Model	#FLOPs	Ratio	#Params	Ratio
Ours	3.42M	1x	3.16M	1x
EfficientNet-B1	0.70B	204.7x	7.80M	2.5x
Inception-v3	5.70B	1667x	24.0M	7.6x
VGG-16	15.3B	4474x	138M	44x
ResNet-34	3.60B	1053x	21.8M	6.9x

machine learning-based methods can deliver performance comparable to that of deep learning-based methods, yet with markedly higher efficiency.

5. Conclusion

In this paper, we introduced a lightweight and generalizable evaluation method, applicable to both efficient machine learning and deep learning contexts. This versatile approach enables sample-level quality assessment, assigning individual quality scores to each generated sample. Additionally, it enhances the performance of generative models by filtering out poor generations. We employ three evaluation metrics to ensure a thorough assessment of both the generative models and individual samples they generate, including accuracy, quality, and diversity. Extensive experiments demonstrate that our method achieves performance consistent with the Fréchet Inception Distance (FID), but at a significantly reduced computational cost and model size.

References

[1] Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning*, pages 290–306. PMLR, 2022. 1

[2] Zohreh Azizi, Xuejing Lei, and C-C Jay Kuo. Noise-aware texture-preserving low-light enhancement. In *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 443–446. IEEE, 2020. 3

[3] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6228–6237, 2018. 2

[4] Ali Borji. Pros and cons of gan evaluation measures. *Computer Vision and Image Understanding*, 179:41–65, 2019. 1

[5] Ali Borji. Pros and cons of gan evaluation measures: New developments. *Computer Vision and Image Understanding*, 215:103329, 2022. 1

[6] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016. 5

[7] Yueru Chen, Mozhdeh Rouhsedaghat, Suya You, Raghuveer Rao, and C-C Jay Kuo. Pixelhop++: A small successive-subspace-learning-based (ssl-based) model for image classification. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 3294–3298. IEEE, 2020. 3

[8] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011. 5

[9] Shuyang Gu, Jianmin Bao, Dong Chen, and Fang Wen. Giqa: Generated image quality assessment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 369–385. Springer, 2020. 2

[10] Swaminathan Gurumurthy, Ravi Kiran Sarvadevabhatla, and R Venkatesh Babu. Deligan: Generative adversarial networks for diverse and limited data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 166–174, 2017. 2

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1, 2, 5

[12] Daniel Jiwoong Im, He Ma, Graham Taylor, and Kristin Branson. Quantitatively evaluating gans with divergences proposed for training. *arXiv preprint arXiv:1803.01045*, 2018. 1, 2

[13] Pranav Kadam, Min Zhang, Shan Liu, and C-C Jay Kuo. Unsupervised point cloud registration via salient points analysis (spa). In *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 5–8. IEEE, 2020. 3

[14] Pranav Kadam, Min Zhang, Shan Liu, and C-C Jay Kuo. R-pointhop: A green, accurate and unsupervised point cloud registration method. *arXiv preprint arXiv:2103.08129*, 2021. 3

[15] Pranav Kadam, Qingyang Zhou, Shan Liu, and C-C Jay Kuo. Pcrp: Unsupervised point cloud object retrieval and pose estimation. *arXiv preprint arXiv:2202.07843*, 2022. 3

[16] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 5

[17] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020. 5

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 5

[19] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[20] C-C Jay Kuo and Azad M Madni. Green learning: Introduction, examples and outlook. *arXiv preprint arXiv:2210.00965*, 2022. 2

[21] C-C Jay Kuo, Min Zhang, Siyang Li, Jiali Duan, and Yueru Chen. Interpretable convolutional neural networks via feedforward design. *Journal of Visual Communication and Image Representation*, 60:346–359, 2019. 3

[22] Xuejing Lei, Ganning Zhao, and C.-C. Jay Kuo. NITES: A non-parametric interpretable texture synthesis method. In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1698–1706. IEEE, 2020. 3

[23] Xuejing Lei, Ganning Zhao, Kaitai Zhang, and C-C Jay Kuo. Tghop: an explainable, efficient, and lightweight method for texture generation. *APSIPA Transactions on Signal and Information Processing*, 10, 2021. 3

[24] Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. Agiqa-3k: An open database for ai-generated image quality assessment. *arXiv preprint arXiv:2306.04717*, 2023. 2

[25] Shaohui Liu, Yi Wei, Jiwen Lu, and Jie Zhou. An improved evaluation framework for generative adversarial networks. *arXiv preprint arXiv:1803.07474*, 2018. 2

[26] Shan Liu, Min Zhang, Pranav Kadam, and Chung-Chieh Jay Kuo. *3D Point Cloud Analysis: Traditional, Deep Learning, and Explainable Machine Learning Methods*. Springer. 3

[27] David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. *arXiv preprint arXiv:1610.06545*, 2016. 1, 2, 4

[28] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. *Advances in neural information processing systems*, 31, 2018. 2

[29] Zhanxuan Mei, Yun-Cheng Wang, Xingze He, and C-C Jay Kuo. Greenbiqa: A lightweight blind image quality assessment method. *arXiv preprint arXiv:2206.14400*, 2022. 3

[30] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, pages 7176–7185. PMLR, 2020. 1, 2

[31] Jeeseung Park and Younggeun Kim. Styleformer: Transformer based generative adversarial networks with style vector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8983–8992, 2022. 5

[32] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018. 1

[33] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 1, 2, 5

[34] Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. *Advances in Neural Information Processing Systems*, 34:17480–17492, 2021. 5

[35] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings*, pages 1–10, 2022. 5

[36] Tingwei Shen, Ganning Zhao, and Suya You. A study on improving realism of synthetic data for machine learning. In *Synthetic Data for Artificial Intelligence and Machine Learning: Tools, Techniques, and Applications*, volume 12529, pages 251–258. SPIE, 2023. 1

[37] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 5

[38] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8):3998–4011, 2018. 2

[39] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 5

[40] Yuan Tian, Qin Wang, Zhiwu Huang, Wen Li, Dengxin Dai, Minghao Yang, Jun Wang, and Olga Fink. Off-policy reinforcement learning for efficient and effective gan architecture search. In *European Conference on Computer Vision*, pages 175–192. Springer, 2020. 5

[41] Zhendong Wang, Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Diffusion-gan: Training gans with diffusion. *arXiv preprint arXiv:2206.02262*, 2022. 5

[42] Tian Xie, Rajgopal Kannan, and C-C Jay Kuo. Graphhop++: New insights into graphhop and its enhancement. *arXiv preprint arXiv:2204.08646*, 2022. 3

[43] Tian Xie, Bin Wang, and C-C Jay Kuo. Graphhop: An enhanced label propagation method for node classification. *arXiv preprint arXiv:2101.02326*, 2021. 3

[44] Yijing Yang, Vasileios Magoulianitis, and C-C Jay Kuo. E-pixelhop: An enhanced pixelhop method for object classification. In *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1475–1482. IEEE, 2021. 2

[45] Yijing Yang, Wei Wang, Hongyu Fu, and C-C Jay Kuo. On supervised feature selection from high dimensional feature spaces. *arXiv preprint arXiv:2203.11924*, 2022. 3

[46] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 5

[47] Min Zhang, Pranav Kadam, Shan Liu, and C-C Jay Kuo. Gsip: Green semantic segmentation of large-scale indoor point clouds. *arXiv preprint arXiv:2109.11835*, 2021. 3

[48] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 3

[49] Xinfeng Zhang, Sam Kwong, and C-C Jay Kuo. Data-driven transform-based compressed image quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(9):3352–3365, 2020. 3

[50] Ganning Zhao, Wenhui Cui, Suya You, and C-C Jay Kuo. Semst: Semantically consistent multi-scale image translation via structure-texture alignment. *arXiv preprint arXiv:2310.04995*, 2023. 1

[51] Ganning Zhao, Vasileios Magoulianitis, Suya You, and C-C Jay Kuo. Lgsqe: Lightweight generated sample quality evaluation. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 1915–1919. IEEE, 2023. 1

[52] Ganning Zhao, Vasileios Magoulianitis, Suya You, C-C Jay Kuo, et al. Lightweight quality evaluation of generated samples and generative models. *APSIPA Transactions on Signal and Information Processing*, 12(1). 1, 5

[53] Ganning Zhao, Tingwei Shen, Suya You, and C-C Jay Kuo. Unsupervised synthetic image refinement via contrastive learning and consistent semantic-structural constraints. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications V*, volume 12538, pages 440–449. SPIE, 2023. 1

[54] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428*, 2021. 1, 2, 4