Diverse Sample Generation: Pushing the Limit of Generative Data-Free Quantization

Haotong Qin, Yifu Ding, Xiangguo Zhang, Jiakai Wang, Xianglong Liu, Member, IEEE, and Jiwen Lu, Senior Member, IEEE

Abstract—Generative data-free quantization emerges as a practical compression approach that quantizes deep neural networks to low bit-width without accessing the real data. This approach generates data utilizing batch normalization (BN) statistics of the full-precision networks to quantize the networks. However, it always faces the serious challenges of accuracy degradation in practice. We first give a theoretical analysis that the diversity of synthetic samples is crucial for the data-free quantization, while in existing approaches, the synthetic data completely constrained by BN statistics experimentally exhibit severe homogenization at distribution and sample levels. This paper presents a generic Diverse Sample Generation (DSG) scheme for the generative data-free quantization, to mitigate detrimental homogenization. We first slack the statistics alignment for features in the BN layer to relax the distribution constraint. Then, we strengthen the loss impact of the specific BN layers for different samples and inhibit the correlation among samples in the generation process, to diversify samples from the statistical and spatial perspectives, respectively. Comprehensive experiments show that for large-scale image classification tasks, our DSG can consistently quantization performance on different neural architectures, especially under ultra-low bit-width. And data diversification caused by our DSG brings a general gain to various quantization-aware training and post-training quantization approaches, demonstrating its generality and effectiveness.

Index Terms—Data-free quantization, deep learning, model compression, quantized neural networks.

I. INTRODUCTION

With the advent of deep learning, the deep neural network has achieved great success in a variety of fields, such as image classification [31], [39], [53], object detection [20], [21], [45], [50], semantic segmentation [17], [66], etc. Nevertheless, it is still a significant challenge to apply advanced neural networks on resource-limited devices for their high memory usage and expensive computation. With more and more hardware supporting low bit-width computations, network quantization emerges as an efficient method to compress and accelerate models [5], [15], [48], [49], [56], [57], [59], [60], [67]. Many quantization methods, called quantization-aware training (QAT), apply the following pipeline: considering the quantization function in the training process on the original dataset, and minimizing the loss caused by the quantization through backward propagation. Since QAT methods require the finetuning steps, it is considered to be time-consuming and computationally intensive [23], [30], [47]. Thus, quantization without training or finetuning process is also demanded in the industry, which is called post-training quantization (PTQ) in recent studies [2], [13], [37], [42], [65]. Most quantization approaches are designed for the data-driven scenario, and thus require real data in their quantization process. However, real data is not always accessible for privacy or security concerns (e.g., medical and user data). Fortunately, data-free quantization is proposed to quantize deep neural networks without accessing real data [7], [34], [35], [43], [52]. As shown in Fig. 1, the data-free quantization approaches can also be classified as data-free PTQ and data-free QAT according to whether the model training is required. Among existing data-free quantization methods, generative methods calibrate or train

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
networks using the “optimal synthetic data”, which with the distribution best matches the Batch Normalization (BN) statistics of the original full-precision neural network. Generative data-free PTQ methods apply BN statistics loss to directly update the generated data and then progress a computation-saving calibration process, while QAT ones generally introduce a separate generator for synthesizing data towards higher accuracy, and train it and the quantized network jointly.

However, among existing quantization approaches, the data-driven methods always achieve significantly higher accuracy compared with data-free ones, whether for PTQ or QAT. One intuitive reason is that the difference between real data and synthetic data leads to the gap in the performance between data-driven and data-free quantization approaches. As a result, many approaches attempt to improve on synthetic data so that the performance of generative data-free quantization methods continues to approach data-driven quantization. In this work, we first give a theoretical analysis from the diversity view for generative data-free quantization. Our analysis shows that synthetic samples should be distributed as dispersed in the input space while satisfying given constraints. This is often described as the diversification of synthetic samples and would facilitate the optimization of the quantization to cover its potential input space as possible. Compared with real training and testing datasets that usually approximately satisfy the Independent and Identically Distributed (IID) assumption, synthetic data generated according to certain constraints in generative methods are difficult to meet the diversity requirement.

Corresponding to our theoretical analysis, we also experimentally reveal that severe homogenization exists in the data generation processes of existing typical data-free generative quantization methods (ZeroQ and GDFQ methods for PTQ and QAT, respectively).

First, the synthetic data is constrained to match the BN statistics, and thereby the feature distribution might overfit the BN statistics in each layer when the data forward propagate in models. As shown in Fig. 2(a) and (b), the distribution of samples synthesized by existing PTQ and QAT methods almost strictly follows the Gaussian distribution with corresponding BN mean and variance, while the distribution of real data has an obvious offset and enjoys a more diverse distribution.

We consider the first phenomenon as the distribution-level homogenization.

Second, all samples of synthetic data are updated by one specific objective function in typical generative data-free quantization methods, all synthetic samples are applied to the same constraint and directly sum the loss term of each layer. In Fig. 2(c) and (d), the distribution statistics of real data are dispersed while the data generated by existing approaches are centralized.

Third, for data-free QAT methods, synthetic samples are synthesized by a generator network. In the process of learning the distributions of synthetic data for quantization, both the generator and the quantized network may converge to a trivial solution where the former learns to produce few modes exclusively, which is referred to as mode collapse and causes the synthetic data to clustered in sample space. As Fig. 2(e) and (f) show, the synthetic data is aggregated while the real data is scattered.

The second and third phenomena are considered to be sample-level homogenization from the statistical and spatial perspectives.

In a word, distribution-level homogenization means that the overall distribution of synthetic data is strictly restricted to the specific distribution with BN statistics, while sample-level homogenization means the little differences among samples from the statistical and spatial perspectives. Only mitigating homogenization at one certain level cannot ensure the diversity of data at the other level.

To alleviate the accuracy degeneration of the quantized neural network caused by the homogenization of synthetic data, this paper presents a generic data generation scheme, Diverse
Sample Generation (DSG), for generative data-free quantization to enhance the diversity of the synthetic data. The proposed DSG scheme mainly relies on three novel techniques: Slack Distribution Alignment (SDA): relax the distribution constraint of synthetic data by slackening the feature statistics alignment in each BN layer. Layerwise Sample Enhancement (LSE): strengthen the impact of statistics loss of the specific BN layer for its corresponding synthetic sample by applying a layerwise enhancement. Sample Correlation Inhibition (SCI): weaken the correlation among the synthetic samples by applying determinantal point processes loss to the intermediate features in the generation process. Among these techniques, SDA diversifies the synthetic data at the distribution level, while LSE and SCI diversify it at the sample level from the statistical and spatial perspectives, respectively. Considering the generality, DSG focuses on the improvement of the generation process while could be almost decoupled from specific quantization methods. Therefore, the proposed techniques of our DSG can be flexibly applied to different PTQ and QAT quantization approaches effectively, improving the accuracy of quantized neural networks by diversifying synthetic data in the generation process.

Our scheme presents a novel perspective of data diversity for generative data-free quantization, and extensive experiments show that DSG significantly improves both data-free PTQ and QAT. The DSG performs remarkably well across several mainstream neural architectures, including VGG16bn [53] (the VGG16 with BN for the dense layers), ResNet18/20/50 [26], SqueezeNext [19], InceptionV3 [55], ShuffleNet [64], and MobileNetV2 [51], and surpasses the existing data-free methods in a wide margin on the large-scale image classification task and achieves the state-of-the-art (SOTA) results. The quantized networks quantized by our DSG even outperform networks quantized by the real data under various settings. The performance of the quantized neural networks trained by our DSG scheme is pushed to that of their full-precision counterparts, meanwhile, the quantization process gets rid of the data dependence.

We summarize the main contributions of this paper as:

- We first give a theoretical analysis of the utility of data diversity in generative data-free quantization. Our analysis shows that synthetic samples should be distributed as dispersed while satisfying given constraints, which would facilitate the optimization of quantization to cover its potential input space as possible.
- We experimentally show and analyze the homogenization of synthetic data in existing generative data-free quantization. Our study presents the homogenization at the distribution and sample levels and reveals that sample-level homogenization is not limited to the statistical perspective, but also the spatial perspective.
- We propose a novel generic DSG scheme for generative data-free quantization from a comprehensive perspective of data diversity, which effectively improves the data-free PTQ and QAT. DSG presents a novel Sample Correlation Inhibition (SCI), in conjunction with the Slack Distribution Alignment (SDA) and Layerwise Sample Enhancement (LSE) techniques to diversify synthetic data at distribution and sample levels.
- We conduct a detailed ablation study on the proposed DSG, which presents the effectiveness of the proposed techniques (SDA, LSE, and SCI) in generative data-free PTQ and QAT. And the comprehensive evaluation of the DSG scheme shows that our DSG surpasses the existing SOTA methods by a wide margin on various neural architectures and bit-widths, which demonstrates that diversity is an important property of high-quality synthetic data.

Note that our paper extends the preliminary conference paper [63]. This manuscript first gives a theoretical analysis that diversifying the synthetic samples is a crucial element for improving the data-free quantization process, and experimentally reveals that the sample-level homogenization is not limited to the statistical perspective, but also the spatial perspective. This work proposed a generic DSG scheme for generative data-free quantization, including both PTQ and QAT, while the previous scheme proposed in the conference paper can be regarded as a special case for PTQ. The DSG in this work further presents a novel SCI technique to tackle sample-level homogenization from the spatial perspective, which can both significantly improves the performance of various generative data-free quantization approaches. We evaluate our DSG scheme on various neural architectures and compared it with more generative data-free quantization methods. Besides, we further add more analysis and discussion of synthetic data and evaluate the data with various methods. The results present that diversity is an important property of high-quality synthetic data in generative data-free quantization.

II. RELATED WORK

Data-Driven Quantization. As a widely used compression technique, quantization earns lots of attention in recent years. To mitigate the severe accuracy drop resulting from quantization, many studies, such as [23], [30], [60], [67], utilize quantization-aware training methods to regain the accuracy. Generally, they can always provide performance improvements. However, the training process is quite cumbersome for huge computational and time costs. Another critical deficiency is that the datasets are not always ready to use, especially for privacy and security concerns. Without any finetuning or training process, research on post-training quantization has been conducted to improve the trade-off between accuracy and computational cost. [2] introduces two closed-form analytical solutions targeting clipping approximation and per-channel bit allocation combined with bias correction to push the limit of weights and activations quantization to 4-bit. [13] analyzes the mean squared error of quantization and minimizes it via the OMSE method. Unlike previous work which aims to find an appropriate clipping range to deal with outliers in the bell-shaped distribution, [65] proposes outlier channel splitting while remaining a functionally identical network. Not restricted to the round-to-nearest method, [42] exploits a more flexible weight-rounding mechanism called AdaRound, which optimizes task loss due to quantization w.r.t. the preactivation. However, these methods still require access to limited data for a more resilient and better-quantized model.

Data-Free Quantization. Several recent studies have explored quantization methods without training and even calibration
datasets. [43] uses weight equalization to acquire similar weight ranges among channels and bias correction for activations. However, quantizing models to run in 6-bit or lower bit-width is a non-trivial task. Works like [24], [6], and [28] leverage BN statistics to reconstruct "realistic" data. Both [24] and [52] utilize the knowledge distillation scheme to finetune the model, while it might cost a little more time. [24] proposes an inception scheme to generate data, while [28] utilizes generative modeling to deal with the distribution of BN statistics. Besides, [6] can work on multi-bit quantization utilizing the Pareto frontier. [11] and [40] focuses on adversarial approaches. [11] proposes a data-free adversarial knowledge distillation method, which minimizes the maximum distance between the outputs of the full-precision teacher model and the quantized student model for any adversarial samples from a generator. [40] instead not only considers the output of the model layer but also maximizes the discrepancy of feature map in intermediate inter-channel between teacher and student model by the proposed Channel Relation Map. Although the above methods use distinct tricks to synthesize data and improve model accuracy, they all have similar limitations and we take [6] as an example to illustrate in Section III-C specifically.

III. METHODS

In this section, we first revisit the previous works of generative data-free quantization. Then, we reveal the data homogenization phenomenon in existing methods and improve the generation process to diversify the synthetic samples.

A. Preliminaries

Existing data-driven quantization methods require original training and testing datasets to calibrate or finetune the quantized network for higher performance. But the real data is not accessible in many practical applications for privacy or security concerns, and thereby general quantization schemes cannot be applied directly to these scenarios. The data-free quantization is proposed to quantize the network without access to real data.

Since the real data is inaccessible, the knowledge of the real data in the pre-trained network should be fully exploited in data-free quantization. The statistics (mean and standard deviation) of BN layers in full-precision models fit the original real dataset in the training process. Therefore, most data-free quantization schemes construct the synthetic data by using BN statistics loss to utilize the information in the BN layers. The following optimization objective enables the distribution of synthetic data $\mathbf{x}^s$ to best match the BN statistics, including the mean and standard deviation:

$$\min_{\mathbf{x}^s} \mathcal{L}_{\text{BN}} = \sum_{i=1}^{N} \left( \| \tilde{\mu}_i^s - \mu_i \|_2^2 + \| \tilde{\sigma}_i^s - \sigma_i \|_2^2 \right),$$  \quad (1)$$

where $\tilde{\mu}_i^s$ and $\tilde{\sigma}_i^s$ are the mean and standard deviation of feature distribution of synthetic data $\mathbf{x}^s$ at the $i$th BN layer, while $\mu_i$ and $\sigma_i$ are that of pre-trained full-precision model $\mathbf{M}$. By minimizing the BN statistics loss $\mathcal{L}_{\text{BN}}$, the synthetic data or the generator learns a distribution of input data to best match the BN statistics of each layer.

Among most of the mainstream generative data-free quantization methods, the PTQ methods directly update the data from random vectors during the generation process and then use the synthetic data to calibrate the quantized neural network. The typical optimization objective in the generation process of these methods can be expressed as

$$\min_{\mathbf{x}^s} \mathcal{L}_{\text{PTQ}} (\mathbf{x}^s, \mathbf{M}),$$  \quad (2)$$

where $\mathbf{M}$ denotes the full-precision model, and the loss function $\mathcal{L}_{\text{PTQ}}$ is usually the BN statistic loss $\mathcal{L}_{\text{BN}}$. The synthetic data $\mathbf{x}^s$ is then used to calibrate the quantized model, a process that is usually completely separate from generation.

While the generative data-free QAT methods train a separate generator model to generate the synthetic data and use them to train the quantized neural network. The typical generation process of these methods is expressed as follows:

$$\min_{\mathbf{G}} \mathcal{L}_{\text{QAT}} (\mathbf{G} (\mathbf{x}^s), \mathbf{M}), \quad \mathbf{x}^s = \mathbf{G} (\mathbf{x}^s^*),$$  \quad (3)$$

where $\mathbf{x}^s^*$ is an initial tensor drawn from standard Gaussian distribution $\mathcal{N}(0, 1)$, the loss function $\mathcal{L}_{\text{QAT}}$ usually consists of the BN statistical loss and the cross-entropy function of the generator $\mathbf{G}$, etc., and $\mathbf{x}^s$ and $\mathbf{y}$ are synthetic data and the corresponding label, respectively. And the quantized neural network $\mathbf{Q}$ is trained simultaneously using synthetic data:

$$\min_{\mathbf{Q}} \mathbb{E}_{\mathbf{x}^s, \mathbf{y}} [\mathcal{C} (\mathbf{Q}, (\mathbf{x}^s; \mathbf{y}))],$$  \quad (4)$$

where $\mathcal{C} (\cdot, \cdot)$ is a loss function for quantized neural network $\mathbf{Q}$, such as cross-entropy loss and mean squared error, and $\mathbf{y}$ is the corresponding labels of synthetic data $\mathbf{x}^s$.

B. Diversity of Synthetic Data for Quantization

In generative data-free quantization and other related generative works, diverse synthetic data is always more welcome intuitively. But there is little relevant theoretical analysis to support this intuition, especially in the context of data synthesis with strict constraints. It is important to understand the theoretical motivation behind this intuition. Therefore, we give a theoretical analysis to show that in various generative data-free quantization approaches, the diversity of synthetic data facilitates the quantization process.

**Definition 1.** For a well-trained full-precision model $\mathbf{M}$, the optimization objective of the whole generative data-free quantization process can be abstracted as

$$\min_{\mathbf{x}^s, \mathbf{Q}} \mathcal{L}_1 (\mathbf{x}^s, \mathbf{M}) + \mathcal{L}_2 (\mathbf{Q}, \mathbf{x}^s), \quad \mathbf{x}^s \subseteq \mathcal{X},$$

where $\mathbf{x}^s$ is the synthetic data, $\mathbf{M}$ and $\mathbf{Q}$ are the full-precision and quantized networks, respectively, and $\mathcal{X}$ denotes the input space of their networks. $\mathcal{L}_1$ and $\mathcal{L}_2$ are constraints for the data generation and network quantization, respectively.

According to the discussion in Section III-A, we can summarize that, in generative data-free quantization methods, the
key elements are synthetic data, full-precision model, and corresponding quantized model. The full-precision model is well-trained and fixed, while the quantized model and synthetic data are optimizable. Therefore, we abstract and define the general generative data-free quantization process as Definition 1. In this summarized definition, a crucial implicit premise is that the synthetic data $x^s$ should reflect the input space $X$ in the process. The optimization process of the quantized model $Q$ will perform well only if this premise is well satisfied. However, the input space $X$ is not explicit in the process, and the real data, which is usually regarded as the ideal set reflecting the input space, cannot be obtained.

**Lemma 1.** Under the ideal low-density and smoothness assumptions, and given the definitions of high-density and low-high regions: (a) any two high-density regions are disconnected; (b) the density of any region connected to a high-density region is lower than that of the high-density region; (c) the union of high-density and low-density regions is equal to the entire input space. Let $X$ be an input space with multiple classes, where the number of samples is much greater than the number of classes. Then, there exists a partition of $X$ into high-density regions $\{R_{H1}, \ldots, R_{Hk}\}$ and low-density regions $\{R_{L1}, \ldots, R_{Ll}\}$, where $h \geq 2$ and $l \geq 1$.

Therefore, for generative data-free quantization, it is necessary to study how to allow synthetic samples to reflect the properties of an unknown input space well. Fortunately, there are two classical assumptions about the input space concluding multiple classes of data in deep learning, and they help reveal the properties of the ideal set of synthetic samples. (1) **Low-density assumption.** The assumption indicates that the boundary of class regions the decision boundary should not pass through high-density regions [3], [58]. (2) **Smoothness assumption.** The assumption states that for any two input points that are close by in the input space, the corresponding labels should be the same [8], [58]. As the Lemma 1, these assumptions allow us to model the potential input space as several independent sub-regions, and we provide the proof of this modeling in Appendix A.1, available online. When we model a potential input space $X$ in generative data-free quantization in this way, consider a discrete set of synthetic samples $x^s = \{x^s_0, x^s_1, \ldots, x^s_N\}$, we can intuitively get that a necessary condition for the sample set to well reflect the input space is to well reflect all sub-regions of the input space because each sub-regions are divided as independent and are not representative for others. And for the sample generation, the specific shape and characteristics of the potential input space are uncertain, so the discussion of the input space should face all possible input spaces rather than a specific one.

**Theorem 1.** Given a set of all possible input spaces $X = \{X_1, X_2, \cdots\}$, where each element $X_i$ has a corresponding overall density $\rho_i$, volume $V_i$, and mass $M_i = \rho_i \cdot V_i$. Additionally, $X_i$ is comprised of several regions $\{R_{i1}, \ldots, R_{iK_i}\}$ with unknown, limited $K_i \geq 2$. The region $R_{ij}$, $1 \leq j \leq K_i$ also have its own density $\rho_j$, volume $V_j$, and mass $M_j$. Let us consider a sample set $x^s = \{x^s_1, \ldots, x^s_N\} \subset X^s$, where $X^s = \mathbb{E}(X)$ represents the latent input space with mass $M^s$. And $X^* = \bigcup_{j=1}^{K} R^*_{j}$, where the mass of $R^*_{j}$ is $M^*_j$ and $K^* = \max(K^1, K^2, \cdots)$. If $x^s$ satisfies the condition that $\forall x^s_i \in x^s$, $p(x^s_i \in R^*_j) = \frac{M^*_j}{M^s}$, the information reflected from $X$ by the sample set $x^s$ will be maximized in mathematical expectation, where $M^s = \sum_{j=1}^{K^*} M^*_j$.

**Theorem 1** shows that the optimized strategy to make synthetic samples adequately reflect the unknown input space in expectation is to correlate the distribution of samples with the spatial scales of the sub-regions, i.e., dispersing synthetic samples uniformly in the potential input space instead of gathering collapse. We present the proof of the theorem in Appendix A.2, available online. In practice, this means that synthetic samples should be distributed as dispersed as possible in the space while satisfying given constraints, which is often described as the diversification of synthetic samples.

### C. Homogenization of Synthetic Data

The generative data-free quantization well sidesteps the issue of lacking accessibility of real data, but we have to concede the huge performance drop compared with networks quantized with real data. As the theoretical analysis in Section III-B, we show that the diversity of synthetic data is beneficial for generative data-free quantization, thus we attempt to find the performance bottleneck of existing methods from this aspect. By experimental analysis for the typical generative data-free PTQ and QAT methods with the almost independent generation process (ZeroQ and GDFQ, respectively), we discover the homogenization phenomena from the existing image synthesizing methods in data-free quantization, which degrade the fidelity and quality of synthetic data. We conclude the homogenization of synthetic data into the distribution level and the sample level to elaborate on them.

1. **Distribution-Level Homogenization:** Since the performance of existing data-driven quantization significantly surpasses that of data-free quantization in usual, the original training data (real data) is generally considered as the upper limit of the performance of synthetic data. So as (1) shows, methods constrained by $\mathcal{L}_{BN}$ are dedicated to generating samples by confining the feature statistics of synthetic data around BN statistics of the pre-trained model. However, the data from the real world is always more diverse, while that generated by fitting BN statistics suffers constrained features as shown in Fig. 2(a) and (b). The activation distribution of data generated in ZeroQ and GDFQ well fits the normal distribution of BN statistics while that of real data deviates from BN statistics. When we calculate the Wasserstein distance between the Gaussian distribution with BN statistics and the feature distribution of synthetic data, the index values of the existing methods are significantly smaller than the real data (real data 0.120 vs ZeroQ 0.040 in PTQ, and real data 0.134 vs GDFQ 0.116 in QAT). We consider it as distribution-level homogenization caused by the strict constraint applied to the synthesizing process makes the synthetic data overfit BN statistics in each layer.

2. **Sample-Level Homogenization:** Besides homogenization at the distribution level, synthetic data in existing generative data-free quantization also suffer homogenization at the sample level.
From the Statistical perspective, in existing generative data-free methods, such as ZeroQ and GDFQ, all samples are constrained by the identical form of BN statistics loss. So there might be little difference among statistics of synthetic samples in one batch, while samples from the real world are diverse. As shown in Fig. 2(c) and (d), the points of the mean and standard deviation of ZeroQ and GDFQ data are significantly overlapping and all points gather together closely near BN statistics. However, statistics of real data and our DSG data have larger variances and are more dispersed. Taking PTQ as an example, the variance of the statistics of the real sample is 0.029, that much larger than the variance of data generated by ZeroQ, which is 0.009.

From the Spatial perspective, the synthetic samples generated in different ways have different behaviors. For existing data-free methods, all samples are synthesized by the same generation strategy. The samples generated by GDFQ are concentrated in spatial perspective, as shown in Fig. 2(f), while samples of our DSG and real data are more scattered. The heatmap shows the degree of aggregation in the synthetic data sample space. The densest cluster of GDFQ samples appears red in the heatmap and the density index is up to 2052 near the cluster center, while the highest density index of the real sample is only 1681. For PTQ, the index is 1902, and its homogenization phenomenon is not significant as QAT but is more severe than the real data. We regard these as sample-level homogenization from the statistical and spatial perspectives, which leads to a significant performance drop for the networks quantized by synthetic data.

In a nutshell, for the generation process in data-free quantization, the homogenization exists in two levels, i.e., the distribution and sample levels. Thus the models quantized by these homogenized data can hardly achieve high accuracy. In this work, we devote to diversify the synthetic data to improve the generative data-free quantization. We propose a generic data-free quantization scheme to tackle the homogenization problems by diversifying synthetic data from different perspectives, which we call Diverse Sample Generation (DSG) scheme. The synthetic data generated by our scheme enables the quantized neural network to achieve high accuracy.

D. Diverse Sample Generation

We present a generic Diverse Sample Generation (DSG) scheme for both generative data-free PTQ and QAT approaches (Fig. 3), including three novel techniques to alleviate homogenization: Slack Distribution Alignment in Section III-D1 is to relax the statistical constraint and thus mitigate distribution-level homogenization, while Layerwise Sample Enhancement in Section III-D2 and Sample Correlation Inhibition in Section II-L-D3 diversify the synthetic data at the sample level from distribution and spatial perspectives, respectively.

1) Slack Distribution Alignment: To deal with the distribution-level homogenization, we propose Slack Distribution Alignment (SDA). Specifically, we apply the relaxation constants during the generation process for matching original BN statistics, which can be intuitively regarded as margins for means and standard deviation. With the relaxation constants, the distribution statistics of synthetic samples do not need to fit the BN statistics strictly, and thus the data can be more diverse in distribution. The loss term of SDA is

\[
l_{SDA} = \| \max (|\tilde{\mu}_i - \mu_i| - \delta_i, 0) \|_2^2 + \| \max (|\tilde{\sigma}_i - \sigma_i| - \gamma_i, 0) \|_2^2,
\]

where \(\delta_i\) and \(\gamma_i\) denote the relaxation constants for the mean and standard deviation statistics of features at the \(i\)th BN layer, respectively. After introducing relaxation to the constraint between statistics of synthetic data and BN layers, the mean and standard deviation of generated samples can fluctuate within a certain range. Benefiting from the relaxed constraints, the synthetic samples are more diverse and their feature distribution behaves closer to that of real data. As the ideal data to quantize networks, our experimental evidence in Fig. 2 shows that the feature statistics of real data still exist gaps compared with BN statistics. Thus, we attempt to approximate the gap between the statistics of real data and BN statistics of the original model as a reference to determine the value of \(\delta_i\) and \(\gamma_i\). For real-world
data, when the number of samples achieves a huge amount, the components and features of inputs approximate the Gaussian assumption. Based on the central limit theorem, we suppose the whole potential input set is generally consistent with a Gaussian distribution.

Thus, we set the initial value of $\delta_i$ and $\gamma_i$ by a batch of samples that were randomly initialized from the Gaussian distribution: First, we input 1024 samples initialized by Gaussian distribution with $\mu = 0$ and $\sigma = 1$ into the full-precision network and then save the feature statistics at each BN layer, including the mean and standard deviation of the feature distribution. Then, we measure the margin between saved statistics and the corresponding BN statistics, then take percents of the margin to initialize $\delta_i$ and $\gamma_i$:

$$
\begin{align*}
\delta_i &= |\bar{\mu}_i - \mu_i|_\epsilon, \\
\gamma_i &= |\bar{\sigma}_i - \sigma_i|_\epsilon,
\end{align*}
$$

where $\epsilon \in (0, 1]$ is a hyper-parameter determining the degree of relaxation, and $|\cdot|_\epsilon$ are mean/standard deviation of the feature of Gaussian initialized data $x^0$ at the $i$th BN layer. $|\bar{\mu}_i - \mu_i|_\epsilon$ and $|\bar{\sigma}_i - \sigma_i|_\epsilon$ denote the $\epsilon$ percentile of $|\mu_i^0 - \mu_i|$ and $|\sigma_i^0 - \sigma_i|$, respectively. Intuitively, the larger the value of $\epsilon$, the more relaxing the constraints are in (5). We empirically set the default value of $\epsilon$ as 0.9 to deal with outliers.

2) Layerwise Sample Enhancement: For the generation processes in existing generative data-free quantization approaches, all synthetic samples are always generated by the same initialization and optimization strategy. Specifically, the samples share the same objective function that directly sums the loss of each layer, i.e., all the BN layers receive the same degree of attention from each sample, and thus results in homogenization among the synthetic samples. We propose Layerwise Sample Enhancement (LSE) to tackle sample-level homogenization. LSE enables each sample to focus on the statistics of different layers by enhancing the specific loss term of the sample in the optimization process so that the synthetic samples of our DSG in one batch could be diverse.

Given a well-trained full-precision network containing the $N$ number of BN layers, earlier generative data-free quantization methods treat each layer the same way without particular predilection or bias. But in LSE, we engineer $N$ different loss terms and apply each of them to the specific data sample. Here, we suppose to generate $N$ samples in one batch, equaling the number of BN layers of the quantized network. The loss term of LSE for the $i$th sample is as follows:

$$
l_{\text{LSE},i} = \frac{1}{N} (J_{N,1}^i + h_i^T l),
$$

where $J_{n,m}$ denotes a $n$-row and $m$-column all-one matrix [29], [61], $h_i$ is an $N$-dim one-hot column vector where the 1 is in the $i$th position, $l = [l_1, l_2, \ldots, l_N]^T$ is the vector of the original loss terms for all BN layers. The set of loss functions for all samples can be expressed as $X_{\text{LSE}} = [l_{\text{LSE},1}, l_{\text{LSE},2}, \ldots, l_{\text{LSE},N}]^T$, which is an $N$-dim column vector and its $i$th element represents the loss function for $i$th sample in the batch. Thus, for the whole set of samples $x^s$, the effect exerted by LSE can be formulated as

$$
x_{\text{LSE}} = \frac{1}{N} (J_{N} + [h_1, h_2, \ldots, h_N]),
$$

where $J_{N}$ denotes an $N$-dim all-one square matrix, and $X_{\text{LSE}}$ can be seen as the enhancement matrix for loss terms. In this way, we focus on every BN layer of the original model respectively and generate corresponding samples for each layer. By such layerwise enhancement, we optimize the whole batch of synthetic samples simultaneously.

3) Sample Correlation Inhibition: The above two proposed methods are based on the relaxation of generation constraints, however, the state of samples at the spatial level still cannot be perceived and adjusted directly during the generation process. Therefore, in addition to improving the loss function constraining the statistics of synthetic data, we construct the unsupervised Sample Correlation Inhibition (SCI) loss for the samples to sample-level diversify the synthetic data from the spatial perspective. SCI applies the determinant point process loss to spatially disperse the intermediate features in the generator, thereby inhibiting the correlation among these synthetic samples. The determinant point process $P$ can express negative interactions among samples by the similarity kernel $K_S$ of the elements in a set $S$, where a large value of $K_{ij} ; i,j \in S$ reduces the likelihood of both elements to appear together in a diverse subset [4], [33]. The related works are widely used to generate or select diverse samples in some existing generation tasks for diversifying samples closer to the real data [9], [16], [34], [38]. However, real data is inaccessible in generative data-free quantization, and thereby it cannot be used to construct the constraint of the diversity of synthetic data.

![Fig. 4. The procedures of our DSG for specific quantization approaches. Case (a) is the DSG in PTQ and case (b) is in QAT.](image-url)
Therefore, we utilize random samples in our SCI instead of the real data, and the random data is applied as a lower bound of diversity to alleviate the homogenization of synthetic samples. Inspired by Theorem 1, each dimension of these vectors is initialized by the uniform distribution \( U[0, 1] \) independently and randomly. The vectors initialized in this way are approximately uniform on spatial and fixed during the quantization process, composing a homogenized set that uniformly covers the potential input space. In this way, the homogenization of synthetic data is limited to a level lower than its initial state.

So far we are able to build the complete flow of our SCI. The intermediate features correspond to the data \( x^* \) in the generation process is expressed as \( f = \{ f_1, f_2, \ldots, f_N \} \), where \( f_i \) is the feature of the corresponding synthetic sample. And for the set of features \( f \), we construct a set of noise vectors \( r = \{ r_1, r_2, \ldots, r_N \} \) of the same size. Then we construct the similarity kernels \( K_f \) and \( K_r \) of the features \( f \) and noise vector set \( r \) as the following decomposition proposed in [16]:

\[
K_f = \phi(\mathbf{f})^\top \phi(\mathbf{f}) \quad \text{and} \quad K_r = \phi(\mathbf{r})^\top \phi(\mathbf{r}),
\]

where all \( \phi(f_i) \in \phi(\mathbf{f}) \) and \( \phi(r_i) \in \phi(\mathbf{r}) \) are the \( l_2 \) normalized vectors that guarantees the similarity kernels \( K_f \) and \( K_r \) to be real positive semidefinite matrices. The main idea of our SCI is to learn a similarity kernel \( K_f \) of feature \( f \) and make it less than the similarity kernel \( K_r \) of data \( r \). In this way, we keep the correlation among the features of synthetic samples lower than that among random vectors, thus alleviating the homogenization of synthetic data. Since matching two similarity kernels directly is an unconstrained optimization problem [36], we construct the loss using the major characteristics: eigenvalues and eigenvectors of kernels. Hence, our SCI loss term is defined as follows:

\[
l_{SCI} = \max \left( \sum_i \left( \| \lambda_i^f - \lambda_i^r \|_2 - \lambda_i^f \cos (v_i^f, v_i^r) \right), 0 \right),
\]

where \( \lambda_i^f \) and \( \lambda_i^r \) are the \( i \)th eigenvalues of \( K_f \) and \( K_r \), respectively. \( v_i^f \) and \( v_i^r \) are the eigenvectors, and \( \lambda_i^r \) is the min-max normalized version of the eigenvalues applied to alleviate the effect of noisy structures.

### E. Specialization for Different Quantization Approaches

The motivation of our DSG is to improve the performance of various generative data-free quantization methods by proposing a novel and generic data generation scheme. Therefore, here we derive and present the most typical specializations of DSG for generative data-free PTQ and QAT approaches (Fig. 4).

**For the data-free PTQ approach**, DSG applies SDA to each layer for relaxed BN statistics loss terms, integrates LSE to introduce diverse degrees of attention among different samples, and uses SCI to diversify the distribution of samples in the potential input space. Therefore, the overall loss \( L_{DSG-PTQ} \) can be defined as follows involving the above techniques:

\[
L_{DSG-PTQ} = (X_{LSE}l_{SDA})^T J_{N, 1} + l_{SCI}^T J_{N, 1} + l_{SCI}^T J_{N, 1} + l_{SCI}.
\]

This optimization loss \( L_{DSG-PTQ} \) is used to optimize synthetic data directly. After the generation process is complete, the optimized synthetic data is applied to calibrate the quantized neural network. The calibration process is usually completely separate from the generation process, and the typical calibration methods include [6] and [22]. The DSG for PTQ is elaborated in Algorithm 1, and it is summarized in Fig. 4 as the case (a).

**For the data-free QAT approach**, DSG applies an alternating optimization strategy to update the generator \( G \) and the quantized neural network \( Q \). As shown in (3), we initialize a set of random noise samples and specified labels, feeding it into generator \( G \) to generate the synthetic data and then training the generator \( G \). The loss function \( L_{DSG-QAT} \) for the generator can be expressed as

\[
L_{DSG-QAT} = (X_{LSE}l_{SDA})^T J_{N, 1} + l_{SCI}^T J_{N, 1} + L_{CE}(G),
\]

where \( l_{SCI} = \frac{1}{B} \left[ l_{SCI_1}, l_{SCI_2}, \ldots, l_{SCI_B} \right]^T \) in (11) is the SCI loss terms for intermediate features of each block (\( B \) blocks in total) in generator \( G \). \( L_{CE}(G) \) is the cross-entropy loss. The data synthesized by generator \( G \) is applied to train the quantized neural network according to the usual practice [52], aiming to close the prediction probability distribution of the quantized neural network to that of the full-precision model. In the quantization process, DSG trains the generator \( G \) and the quantized neural network alternately in every epoch, and the pre-trained full-precision network is fixed. The DSG for QAT is elaborated in Algorithm 2 and is further illustrated in Fig. 4 as the case (b).

### Algorithm 1: The Process of DSG Scheme in PTQ.

**Input:** pretrained model \( M \) with \( N \) BN layers, training iterations \( T \).

**Output:** synthetic data: \( x^* \), quantized network \( Q \).

1. Initialize \( x^* \) from Gaussian distribution \( N(0, 1) \);
2. Initialize \( x^* \) from Gaussian distribution \( N(0, 1) \);
3. Get \( \mu_i \) and \( \sigma_i \), from BN layers of \( M \), \( i = 1, 2, \ldots, N \);
4. Forward propagate \( M(x^*) \) and gather activations;
5. Compute \( \delta_i \) and \( \gamma_i \) using Eq. (6);
6. for all \( t = 1, 2, \ldots, T \) do
   - Forward propagate \( M(x^*) \) and gather activations;
   - Get \( \mu_i^* \) and \( \sigma_i^* \) from activations;
   - Compute all \( l_{SDA} \), using Eq. (5);
   - Descend \( L_{DSG-QAT} \) using Eq. (10) and update data \( x^* \);
7. Get synthetic data \( x^* \);
8. Calibrate quantized network \( Q \) using data \( x^* \).

### Algorithm 2: The Process of DSG Scheme in QAT.

**Input:** pretrained model \( M \) with \( N \) BN layers, training iterations \( T \).

**Output:** generator network \( G \), quantized network \( Q \).

1. Initialize \( x^{**} \) from Gaussian distribution \( N(0, 1) \) and specify label \( y \);
2. Get \( \mu_i \) and \( \sigma_i \), from BN layers of \( M \), \( i = 1, 2, \ldots, N \);
3. Forward propagate \( M(x^{**}) \) and gather activations;
4. Compute \( \delta_i \) and \( \gamma_i \) using Eq. (6);
5. for all \( t = 1, 2, \ldots, T \) do
   - Generate data \( (x^*, y) \) and gather activations;
   - Forward propagate \( M(x^*) \) and gather activations;
   - Descend \( L_{DSG-QAT} \) using Eq. (11) and update \( G \);
   - Update the quantized network \( Q \) using data \( (x^*, y) \);
7. Train generator network \( G \) and generate data \( (x^*, y) \);
8. Train quantized network \( Q \) using data \( (x^*, y) \).
F. Analysis and Discussion

The techniques in our DSG scheme, SDA, LSE, and SCI, alleviate the homogenization of synthetic data at the distribution level and sample level. In Fig. 2, we statistically observe the behavior of the three types of data, i.e., data generated by our DSG scheme, data generated by other SOTA generative data-free quantization methods, and real data. Here we analyze and discuss the effect of our DSG scheme in diversifying synthetic data in generative data-free quantization methods.

We first analyze the synthetic data generated by our DSG scheme in PTQ. Compared with the data generated by the existing generative data-free PTQ method (ZeroQ), the behavior of samples generated with our DSG is more similar to that of real data statistically. As shown in Fig. 2(a), the distribution of our DSG data does not strictly fall in close vicinity to BN mean and standard deviation, instead, they are more varied in value and thus more fluctuate in frequency. As introduced in Section III-C, the Wasserstein distance between BN statistics and our DSG synthetic data is 0.124 (comparing with 0.040 of ZeroQ). And in Fig. 2(c), the statistics of the data generated by ZeroQ are centralized, while that of the DSG samples are dispersed. The phenomenon means that our synthetic data significantly alleviates the sample-level homogenization of existing data-free generative PTQ methods from the statistical perspective. As mentioned in Section III-C, we calculate the variance of the statistics of our DSG data and compare it with ZeroQ in PTQ, which is 0.029 versus 0.009. Fig. 2(e) presents the distribution of different types of synthetic data in the sample space. The density index near the cluster center of the synthesized sample generated by DSG is only 1571, which is even lower than that of real data 1681 and far lower than ZeroQ 1902.

For the DSG scheme in QAT, due to the utility of SDA and LSE, the distribution of our DSG data is not constrained strictly and the synthetic samples enjoy more diverse statistics, as Fig. 2(b) and (d) show. The SCI further weakens the correlation among samples by constructing unsupervised loss for intermediate features in the generation process to alleviate the homogenization of the generated samples in sample space. From Fig. 2(f), we observe that the density index of the synthesized sample generated by DSG is far lower than the existing data-free QAT method, which means that the sample-level homogenization is greatly alleviated from the spatial perspective.

Therefore, the data generated by our DSG could be more diverse from various perspectives comprehensively, which may have the potential to be an alternative for different quantization approaches that real data is not accessible.

IV. EXPERIMENTS

In this section, we conduct comprehensive experiments to validate the performance of our DSG scheme on image classification tasks. We first conduct an ablation study to show the effects of different components, including SDA, LSE, and SCI in Section IV-A. Then in Section IV-B, we compare DSQ with SOTA data-free PTQ and QAT methods respectively across various network architectures. We evaluate the DSG on the CIFAR10 [32] and ImageNet (ILSVRC12) [14] datasets in PTQ while on the ImageNet dataset in QAT. Finally, in Section IV-C, we conduct a further study on synthetic data. Specifically, we analyze and discuss the data diversity, and integrate and evaluate our synthetic data with various calibration methods and data-driven quantization methods. The results show that good diversity is an important property of high-quality synthetic data.

DSG Scheme. In the data-free PTQ, the proposed DSG scheme is used for generating synthetic data while the independent calibration processes are as [6] and [22], and the effectiveness is evaluated by measuring the accuracy of quantized models. Unless otherwise specified, the calibration process for DSG in PTQ is the same as [6]. For the DSG scheme in QAT, we train the generator network to synthesize the data and use it to finetune the quantized network. The generator in the DSG scheme in QAT is constructed following ACGAN [44] as in [52].

Network Architectures. We evaluate our DSG scheme in a wide range of network architectures with various bit-width to prove the versatility of our method. We employ VGG16bn [53], ResNet20/18/50 [26], ShuffleNext [19], InceptionV3 [55], ShuffleNet [64], and MobileNetV2 [51] with various bit-widths, including W4A4 (means 4-bit weight and 4-bit activation), W6A6, W8A8, etc.

Implementation Details. The proposed scheme is implemented by PyTorch for the sake of the powerful automatic differentiation mechanism. We adopt Gaussian distribution as initialization for the data generation process in our DSG scheme. In our experiments, we quantize all the layers and the activation is clipped in a layerwise manner. As for hyper-parameter (e.g., the number of iterations to provide synthetic data), we mostly follow the released official implementations or the models and settings clarified in their original paper for a fair comparison [6], [22], [52], [63]. For the training and fine-tuning process of DSG in QAT, we use the SGD optimizer in the experiments, where momentum is 0.9 and weight decay is 1e-4. We train quantized networks and generators for 400 epochs. The learning rates for quantized networks and generators are initialized to 1e-6 and 1e-3, respectively, and both of them are decayed by 0.1 for every 100 epochs.

A. Ablation Study

We investigate the effect of the proposed LSE, SDA, and SCI techniques for our DSG scheme in data-free PTQ and QAT by ablation experiments. We use the ResNet18 architecture with the ImageNet dataset to evaluate our method under the W4A4 b-width setting, which can show the effect of each part more obviously.

1) Effect of SDA: We first analyze the effectiveness of SDA. As discussed earlier, hyper-parameter $\epsilon$ determines the degree of relaxation in SDA. Therefore, in Fig. 5, we verify the impact of different values of $\epsilon$ in the interval (0,1] with moderate step size 0.1, and further add $\epsilon = 0$ to serve as the vanilla case that the fitting of BN statistics is constrained without relaxation. In PTQ Fig. 5(a), when the $\epsilon$ varies from 0 to 0.9, the final performance increases gradually from 26.04% to 33.39%. In QAT Fig. 5(b), the accuracy also presents an overall upward trend (from 60.52% to 61.08%) as the $\epsilon$ varies from 0 to 0.9. However, when $\epsilon = 1$,
the accuracy drops more or less in both scenarios, which is 3.61% in PTQ and 0.02% in QAT. These phenomena forcefully prove that by way of relaxing the constraints, the SDA method brings generated data a certain offset when fitting the BN statistic distribution, which contributes to feature diversity and consequently improves performance. And when $\epsilon$ is set to 1, the degree of slack reaches the limit as all the outliers in $\tilde{\mu}_i^0$ and $\tilde{\sigma}_i^0$ are taken into account, which means the feature distribution of generated data might be out of the reasonable range. Therefore, it might result in better feature diversity in synthetic data but would harm the final accuracy of the quantized network. Therefore, we set the default value of $\epsilon$ as 0.9 empirically to balance divergence and the impact of outliers.

As the results in Table I, the vanilla data-free PTQ method without SDA suffers a severe accuracy degradation by 7.35%, and the SDA also provides 0.56% accuracy promotion in data-free QAT (vanilla 60.52% versus SDA 61.08%). The results reflect that SDA is essential. Compared to the other two parts of our scheme, SDA may provide a major contribution to the final performance.

2) Effect of LSE: The Table I also shows that LSE can improve the performance in both data-free QAT and PTQ. As the ablation results show, in PTQ, the LSE technique gives a non-negligible increment compared with ZeroQ by 1.08%. As for QAT, it also helps the DSG scheme to have a slight improvement compared with the vanilla method by 0.19%.

3) Effect of SCI: Then we evaluate SCI in data-free quantization. The motivation of our SCI is to alleviate sample-level homogenization from the spatial perspective. Table I shows the effects of using SCI. In QAT, compared with the vanilla method, integrating with SCI helps to acquire better accuracy, which is 0.40% higher. Also, in PTQ, the SCI method gains 13.49% accuracy compared with using pure random inputs.

We also conduct a hyper-parameter ablation study for loss terms in Fig. 6. We investigated the effect of the normalized SCI loss and SDA-LSE one (plus CE loss for QAT) in ablation studies by adding weight for SCI loss, i.e., $l_{SCI}^{BNN}$ for PTQ or $l_{SCI}^{BNN,N,1}$ for QAT, and compared in detail 14 ($7 \times 2$) settings, including 0:1, $10^{-2.1}:1$, $10^{-1.1}:1$, 1:1 (ours), 10:1, $10^5:1$, and $10^{5.1}$ in QAT and PTQ. The results show that our default setting wins in both data-free PTQ and QAT. Moreover, we observe that the QAT scheme is more sensitive to the weight of SCI loss, and the PTQ scheme is less sensitive. It suggests that our SCI plays a greater role in the quantization process with generator and model training, preventing homogenization during data synthesis consistently.

Intuitively, these three techniques are motivated by different observations, and also the processes are carefully engineered so hardly interfere with each other. In short, SDA and LSE improve the loss related to BN statistics, aiming to prevent the generated samples from overfitting to BN statistics and making the samples focus on the statistics of different layers. While the SCI for data-free quantization constructs an unsupervised loss to constrain the features in the generator to hold distances among them, which mitigates the sample homogenization. The results show that, in data-free QAT, the DSG scheme equipped with these techniques obtains a 1.66% improvement in total with ResNet18 under the W4A4 setting, which is up to 62.18%. As for data-free PTQ, DSG can also boost the performance to 13.86% with all the techniques.

### Table I

| Method               | No FT | W-bit | A-bit | Top-1  |
|----------------------|-------|-------|-------|--------|
| Baseline             | -     | 32    | 32    | 71.47  |
| Vanilla (ZeroQ)      | ✓     | 4     | 4     | 26.04  |
| Sample Correlation Inhibition | ✓ | 4 | 4 | 39.53 |
| Layerwise Sample Enhancement | ✓ | 4 | 4 | 27.12 |
| Slack Distribution Alignment | ✓ | 4 | 4 | 33.39 |
| DSG (Ours)           | ✓     | 4     | 4     | 39.90  |
| Vanilla (GDFQ)       | ✗     | 4     | 4     | 60.52  |
| Layerwise Sample Enhancement | ✗ | 4 | 4 | 60.71 |
| Sample Correlation Inhibition | ✗ | 4 | 4 | 60.92 |
| Slack Distribution Alignment | ✗ | 4 | 4 | 61.08 |
| DSG (Ours)           | ✗     | 4     | 4     | 62.18  |

Fig. 5. The accuracy comparison of different $\epsilon$ values in (6) on ResNet18 in PTQ and QAT. As $\epsilon$ varies from 0 to 0.9, the final accuracy is mainly on the rise. But a significant drop is caused by the outliers when $\epsilon = 1$.  

Fig. 6. The hyper-parameter ablation study of the weight of SCI loss in PTQ (left) and QAT (right) for 4-bit ResNet18. X-axis denotes the weight applied to the SCI loss and Y-axis shows the gap of each ablation setting compared with our setting.

#### B. Comparison With SOTA Methods

We extensively evaluate our DSG scheme on a wide range of architectures on CIFAR10 and ImageNet datasets for the image classification tasks. We denote the bit-width setting of architectures on CIFAR10 and ImageNet datasets for the quantized network as $W$ w $A$ as $W_aA_b$ where $w$ is the bit-width for weight and $a$ is that for activation, like W8A8, W6A6, W4A4, etc.

1) Comparison With Data-Free PTQ Methods: To evaluate the advantage of the proposed scheme in PTQ, we

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
first compare our DSG against other data-free PTQ methods (ZeroQ [6], DFQ [43], ACIQ [1], MSE [10], KL [54], SQuant [22], OCS [65], and MMD [34])) on CIFAR10 and ImageNet datasets. Among these methods, ZeroQ and SQuant are typical generative data-free PTQ methods, which reconstruct synthetic data and calibrate the quantized network in different ways. Thus, we evaluate the generation performance of our DSG in PTQ with calibration methods from ZeroQ and SQuant, denoted as DSG1 and DSG2. Other quantization methods use weight equalization or analytical clip range to improve the network performance. Besides, we additionally compare our method with OCS, which is also a PTQ method but requires real data for calibration. We evaluate these methods on various bit-width settings, and the results on CIFAR10 and ImageNet datasets are shown as Tables II, III, and IV.

Specifically, on CIFAR10 [32] dataset, we evaluate our DSG with ResNet20 [27] and VGG16bn [53]. The results are shown in Table II. Under all settings on the CIFAR10 dataset, our DSG far exceeds existing SOTA methods. And we highlight that the quantized network calibrated with DSG data completely surpasses the network calibrated with real data in various settings. For example, under the W4A4 settings of ResNet20 and VGG6bn, our method exceeds the calibration with real data by 0.41% and 0.39%, respectively.

As listed in Tables III and IV, the results on ImageNet over various network architectures, including ResNet18/50 [26], SqueezeNext [19], InceptionV3 [55], and ShuffleNet [64], demonstrate that our proposed DSG significantly outperforms previous sample generation methods. As the results show, the accuracy of quantized networks calibrated with our DSG data under the W8A8 setting is almost not decreased and even surpasses the full-precision (W32A32) baseline networks on the ResNet18 and InceptionV3 architectures. The higher accuracy might be attributed to making full use of the potential of the quantized network. Quantizing the network to W8A8 maintains a greater representation capability of the network compared with the lower bit-width quantization (such as W4A4) and brings less quantization error. Thus, under this setting, our diversified data can alleviate the performance degradation of the network quantization while even leading to a better solution compared with the full-precision network. And the advantage of our DSG gets more evident when the bit-width becomes lower. For example, our DSG outperforms ZeroQ on SqueezeNext on the W6A6 setting by more than 26%, and even outperforms real data by 1.54%. And it is noteworthy that in W4A4 cases with ResNet18 architecture, our DSG surpasses ZeroQ by 13.86%, and even surpasses real data by a notable 1.45% with the compared SQuant calibration, which is up to 66.67%. Moreover, with InceptionV3 under W4A4 settings, our DSG helps to acquire 74.02% accuracy, which is higher than SQuant and real data by eminent 0.76% and 0.52%, respectively.

We also replace part of our DSG with other methods and compare them with our original designs. We first replace the L2 loss with KL divergence when learning the distribution from real data. As a way of measuring distances, KL divergence learns distributions more robustly than the L2 norm. However, in this scenario, the KL divergence performs poorly due to the assumptions it is based on. The usage of KL divergence in data-free training requires treating the features of each batch of input data in the BN layer as a Gaussian distribution and matching the Gaussian distribution with BN statistics. However, as shown in Fig. 2(a), the features of the real data are not strictly distributed according to the BN statistic and are even not strictly Gaussian. Therefore, regarding the features of each input batch as Gaussian distribution is actually a strong assumption. And since each batch of data is always initialized from a standard normal distribution in data-free quantization, the optimization of synthetic data with this assumption limits the diversity of features in the generation process. Results using KL divergence are listed in Tables III and IV. It shows that distribution approximation such as KL can improve the calibration, but is lower than the original DSG due to its strong Gaussian assumption. To verify this justification, we add another method KL+(SDA+LSE) for 4-bit and 6-bit quantization on all network architectures. Since our SCI method is proposed to promote the uniform distribution of samples in space, removing the SCI technique in KL+(SDA+LSE) will cause a stronger Gaussian distribution assumption compared with the KL+(SDA+LSE+SCI), rather than the diversity and spatial uniformity emphasized by our DSG. The results of Tables III and IV in our revision show that, on all network architectures, the KL+(SDA+LSE) excluding SCI is worse than KL+(SDA+LSE+SCI). These results further verify our justification and the motivation of DSG. And our methods of SDA, LSE, and SCI still help mitigate homogenization in most cases.

We also replace our SCI with MMD [34] technique, a classic technique that prevents GAN from mode collapses, which is denoted as MMD+(SDA+LSE). And we also tested the effect of using
MMD purely (based on ZeroQ), denoted as MMD. Similar to the DPP in SCI, the MMD-based kernel can also match synthetic data and random data. Results indicate that data synthesized by the MMD technique can improve quantization performance in most 4/6-bit cases, which confirms the motivation of SCI that stretching the distance between samples by applying distance measuring can relieve the homogenization at the sample level, and is beneficial to the diversity of synthetic data. For example, on 4-bit InceptionV3 and ShuffleNet, the MMD outperforms vanilla ZeroQ by 30.48% and 17.90%, respectively. Meanwhile, the results also suggest that MMD can be equipped with SDA and LSE since they target different homogenization problems, and the performance is even better. The results show that our SCI outperforms MMD, implying a reliable improvement bring by custom designs for data-free quantization.

2) Comparison With Data-Free QAT Methods: Furthermore, to demonstrate the applicability of our DSG scheme in data-free QAT, we compare it with existing QAT methods, such as DFC [24], RVQuant [46], Integer-Only [30], ZAQ [40], GZNQ [28], ZS-CGAN [12], and GDFQ [52]. These methods are engineered to finetune and update the network parameters. DFC is a finetuning method to recover accuracy for ultra-low bit-width cases, which uses Inceptionism [41] to facilitate the generation of data with random labels. The other mentioned methods finetune the quantized network to improve the accuracy of the network. Among them, ZAQ and GDFQ introduce a generator to synthesize data and use the generated data to finetune the network. We evaluate these methods on various bit-width settings in the more challenging large-scale image classification task. The results on ImageNet are shown as Table V.

We test these methods on ResNet18/50 [26], InceptionV3 [55], ShuffleNet [64], and MobileNetV2 [51]. The results in Table V show that our DSG enjoys the best performance. Concretely, as can be seen, regardless of network architecture, our DSG method surpasses many previous methods, including DFC, RVQuant, Integer-Only, and GDFQ in various bit-width settings. And it is noteworthy that, after finetuning with our DSG data, quantized ResNet18 and InceptionV3 even surpass their
full-precision counterparts under the W8A8 setting. Compared with other data-free QAT training methods under the W4A4 setting, our DSG outperforms GDFQ by 1.66% and 16.31% with ResNet18 and ResNet50 respectively, also higher than RVQuant with real data by 7.06% with ResNet50. Besides the mainstream neural architectures, we also evaluate our DSG over existing well-designed lightweight architectures. The proposed DSG shows an overwhelming advantage over GDFQ with a similar training pipeline, 9.16% higher with MobileNetV2 under W4A4, 7.93% higher with ShuffleNet under W4a4, and 1.62% higher with InceptionV3 under W4A4.

In short, our DSG scheme outperforms other competitors over a wide range of experiments in data-free PTQ and QAT, including various bit-width, different network architectures, and two datasets. All the results forcefully demonstrate that the diversity of generated data is significant to calibrate the model for higher accuracy, especially in ultra-low bit-width conditions. If the synthetic data are homogeneous or even identical at the distribution and sample level, it would be ineffectual when used to quantize the network. Our DSG effectively diversifies the synthetic data and thus improves the performance of the quantized neural network.

### C. Further Study on Synthetic Data

To further demonstrate the data diversifying caused by our DSG scheme has a general gain to the quantized network, we provide more experiments as corroborations to support our viewpoint. In our paper, we have conducted a bunch of experiments on our synthetic data to evaluate its effectiveness in improving the network performance, including analyzing the data diversity and integrating the synthetic data generated by our method with different calibration methods and data-driven quantization methods. The experiments show that benefits from the increase in the diversity of synthetic data, the network performance in various methods is significantly improved. It shows that good diversity is an important property of high-quality synthetic data.

#### 1) Analysis and Discussion for Data Diversity

In this section, we further discuss the visualization results of our DSG scheme. First, for data-free PTQ, we show the distribution of statistics of the real data, vanilla data (ZeroQ), and DSG data (ours) in Fig. 7, which explains that SDA and LSE alleviate the homogenization from the aforementioned two levels. Then, for data-free QAT, we visualized samples of vanilla data (GDFQ), DSG data (ours), real data, and Gaussian data in Fig. 8 to show

| (a) InceptionV3 | Arch | Method | No D | No FT | W-bit | A-bit | Top-1 |
|-----------------|------|--------|------|-------|-------|-------|------|
| Baseline        | -    | - 32   | 32   | 78.80 |
| Real Data       | x    | v 4    | 4    | 73.50 |
| ZeroQ           | v    | v 4    | 4    | 12.00 |
| KL (SDA+LSE)    | v    | v 4    | 4    | 26.37 |
| KL (SDA+LSE+SC) | v    | v 4    | 4    | 26.42 |
| MMD             | v    | v 4    | 4    | 42.48 |
| KL              | v    | v 4    | 4    | 49.08 |
| MMD (SDA+LSE)   | v    | v 4    | 4    | 57.17 |
| SQuant          | v    | v 4    | 4    | 73.26 |
| DSG1 (Ours)     | v    | v 4    | 4    | 57.17 |
| DSG2 (Ours)     | v    | v 4    | 4    | 74.02 |
| InceptionV3     |      |        |      |       |
| Real Data       | x    | v 6    | 6    | 78.59 |
| ZeroQ           | v    | v 6    | 6    | 74.55 |
| KL (SDA+LSE)    | v    | v 6    | 6    | 75.14 |
| KL (SDA+LSE+SC) | v    | v 6    | 6    | 76.24 |
| MMD             | v    | v 6    | 6    | 77.36 |
| KL              | v    | v 6    | 6    | 77.80 |
| MMD (SDA+LSE)   | v    | v 6    | 6    | 78.18 |
| SQuant          | v    | v 6    | 6    | 78.30 |
| DSG1 (Ours)     | v    | v 6    | 6    | 78.12 |
| DSG2 (Ours)     | v    | v 6    | 6    | 78.59 |
| ShuffleNet      |      |        |      |       |
| Real Data       | x    | v 8    | 8    | 78.79 |
| MMD             | v    | v 8    | 8    | 78.62 |
| ZeroQ           | v    | v 8    | 8    | 78.70 |
| SQuant          | v    | v 8    | 8    | 78.79 |
| MMD (SDA+LSE)   | v    | v 8    | 8    | 78.80 |
| DSG1 (Ours)     | v    | v 8    | 8    | 78.81 |
| DSG2 (Ours)     | v    | v 8    | 8    | 78.85 |

| (b) SqueezeNext and ShuffleNet | Arch | Method | No D | No FT | W-bit | A-bit | Top-1 |
|---------------------------------|------|--------|------|-------|-------|-------|------|
| Baseline                        | -    | - 32   | 32   | 69.36 |
| Real Data                       | x    | v 6    | 6    | 66.51 |
| MMD                             | v    | v 6    | 6    | 37.29 |
| ZeroQ                           | v    | v 6    | 6    | 39.83 |
| KL                              | v    | v 6    | 6    | 64.16 |
| KL (SDA+LSE)                    | v    | v 6    | 6    | 65.08 |
| KL (SDA+LSE+SC)                 | v    | v 6    | 6    | 65.42 |
| MMD (SDA+LSE)                   | v    | v 6    | 6    | 66.19 |
| SQquant                         | v    | v 6    | 6    | 67.34 |
| DSG1 (Ours)                     | v    | v 6    | 6    | 66.23 |
| DSG2 (Ours)                     | v    | v 6    | 6    | 68.05 |
| SqueezeNext                     |      |        |      |       |
| Real Data                       | x    | v 8    | 8    | 69.23 |
| MMD                             | v    | v 8    | 8    | 61.85 |
| ZeroQ                           | v    | v 8    | 8    | 68.01 |
| SQquant                         | v    | v 8    | 8    | 69.22 |
| MMD (SDA+LSE)                   | v    | v 8    | 8    | 69.25 |
| DSG1 (Ours)                     | v    | v 8    | 8    | 69.27 |
| DSG2 (Ours)                     | v    | v 8    | 8    | 69.37 |
| ShuffleNet                      |      |        |      |       |
| Real Data                       | x    | v 8    | 8    | 56.25 |
| MMD                             | v    | v 8    | 8    | 61.65 |
| ZeroQ                           | v    | v 8    | 8    | 39.03 |
| KL                              | v    | v 8    | 8    | 39.92 |
| KL (SDA+LSE)                    | v    | v 8    | 8    | 54.37 |
| MMD                             | v    | v 8    | 8    | 56.93 |
| SQquant                         | v    | v 8    | 8    | 60.25 |
| MMD (SDA+LSE)                   | v    | v 8    | 8    | 60.39 |
| DSG1 (Ours)                     | v    | v 8    | 8    | 60.71 |
| DSG2 (Ours)                     | v    | v 8    | 8    | 61.94 |
| ShuffleNet                      |      |        |      |       |
| Real Data                       | x    | v 8    | 8    | 64.52 |
| MMD                             | v    | v 8    | 8    | 63.95 |
| ZeroQ                           | v    | v 8    | 8    | 64.46 |
| SQquant                         | v    | v 8    | 8    | 64.68 |
| MMD (SDA+LSE)                   | v    | v 8    | 8    | 64.84 |
| DSG1 (Ours)                     | v    | v 8    | 8    | 64.87 |
| DSG2 (Ours)                     | v    | v 8    | 8    | 64.97 |
TABLE V
RESULTS OF DATA-FREE QAT METHODS WITH ResNet18, ResNet50, InceptionV3, ShuffleNet, and MobileNetV2 on ImageNet

(a) ResNet18 and ResNet50

| Arch   | Method       | No D | No FT | W-bit | A-bit | Top-1 |
|--------|--------------|------|-------|-------|-------|-------|
| Baseline | –             | –    | 32    | 32    | 71.74 |
| Real Data | X             | X    | 4     | 4     | 63.87 |
| GDFQ    | ✓             | ✓    | 4     | 4     | 60.52 |
| DSG (Ours) | ✓             | ✓    | 4     | 4     | 62.18 |
| Real Data | X             | X    | 6     | 6     | 71.42 |
| Integer-Only | X             | X    | 6     | 6     | 67.30 |
| GDFQ    | ✓             | ✓    | 6     | 6     | 70.43 |
| DSG (Ours) | ✓             | ✓    | 6     | 6     | 71.12 |
| Real Data | X             | X    | 8     | 8     | 71.44 |
| RVQuant | X             | X    | 8     | 8     | 70.01 |
| DFC     | ✓             | ✓    | 8     | 8     | 69.57 |
| GDFQ    | ✓             | ✓    | 8     | 8     | 71.43 |
| DSG (Ours) | ✓             | ✓    | 8     | 8     | 71.54 |
| Baseline | –             | –    | 32    | 32    | 77.72 |
| Real Data | X             | X    | 4     | 4     | 70.27 |
| GDFQ    | ✓             | ✓    | 4     | 4     | 55.65 |
| RVQuant | X             | X    | 4     | 4     | 64.90 |
| ZAQ     | ✓             | ✓    | 4     | 4     | 70.06 |
| DSG (Ours) | ✓             | ✓    | 4     | 4     | 71.96 |
| ResNet50 | Real Data   | X    | 6     | 6     | 77.56 |
| GDFQ    | ✓             | ✓    | 6     | 6     | 76.59 |
| ZS-CGAN | ✓             | ✓    | 6     | 6     | 76.82 |
| DSG (Ours) | ✓             | ✓    | 6     | 6     | 77.25 |
| Real Data | X             | X    | 8     | 8     | 77.66 |
| GDFQ    | ✓             | ✓    | 8     | 8     | 77.51 |
| DSG (Ours) | ✓             | ✓    | 8     | 8     | 77.64 |

(b) ShuffleNet, MobileNetV2, and InceptionV3

| Arch   | Method       | No D | No FT | W-bit | A-bit | Top-1 |
|--------|--------------|------|-------|-------|-------|-------|
| Baseline | –             | –    | 32    | 32    | 65.07 |
| Real Data | X             | X    | 4     | 4     | 29.18 |
| GDFQ    | ✓             | ✓    | 4     | 4     | 21.78 |
| DSG (Ours) | ✓             | ✓    | 4     | 4     | 29.71 |
| Real Data | X             | X    | 6     | 6     | 62.89 |
| GDFQ    | ✓             | ✓    | 6     | 6     | 60.12 |
| DSG (Ours) | ✓             | ✓    | 6     | 6     | 61.37 |
| Real Data | X             | X    | 8     | 8     | 62.95 |
| GDFQ    | ✓             | ✓    | 8     | 8     | 64.03 |
| DSG (Ours) | ✓             | ✓    | 8     | 8     | 64.76 |
| Baseline | –             | –    | 32    | 32    | 71.88 |
| Real Data | X             | X    | 4     | 4     | 66.39 |
| GDFQ    | ✓             | ✓    | 4     | 4     | 60.50 |
| DSG (Ours) | ✓             | ✓    | 4     | 4     | 60.46 |
| Real Data | X             | X    | 6     | 6     | 72.11 |
| GDFQ    | ✓             | ✓    | 6     | 6     | 70.90 |
| DSG (Ours) | ✓             | ✓    | 6     | 6     | 71.21 |
| Real Data | X             | X    | 8     | 8     | 72.92 |
| GDFQ    | ✓             | ✓    | 8     | 8     | 70.17 |
| DSG (Ours) | ✓             | ✓    | 8     | 8     | 71.43 |
| MobileNetV2 | Integer-Only | X   | 6     | 6     | 72.90 |
| GDFQ    | ✓             | ✓    | 6     | 6     | 71.48 |
| GZQ     | ✓             | ✓    | 6     | 6     | 71.12 |
| DSG (Ours) | ✓             | ✓    | 6     | 6     | 71.48 |
| InceptionV3 | Real Data   | X    | 6     | 6     | 78.80 |
| GDFQ    | ✓             | ✓    | 6     | 6     | 70.39 |
| DSG (Ours) | ✓             | ✓    | 6     | 6     | 72.01 |
| Real Data | X             | X    | 8     | 8     | 78.81 |
| GDFQ    | ✓             | ✓    | 8     | 8     | 77.20 |
| DSG (Ours) | ✓             | ✓    | 8     | 8     | 78.60 |
| Real Data | X             | X    | 8     | 8     | 79.00 |
| GDFQ    | ✓             | ✓    | 8     | 8     | 78.62 |
| DSG (Ours) | ✓             | ✓    | 8     | 8     | 78.94 |

Fig. 7. Mean and standard deviation of the activations in one channel of ResNet18 when feeding different types of data (with 256 samples), including the synthetic data generated by ZeroQ and DSG and the real data. Each sample generated by ZeroQ behaves similarly overfitting BN statistics compared with real data, which shows the homogenization at both distribution and sample levels. Our DSG data enjoys a diversity close to real data to obtain an accurate quantized network.

Fig. 8. Visualization of samples generated by (a) GDFQ method, (b) our SCI method, (c) real data, (d) Gaussian random data. We collect 512 samples for each type of data randomly and reduce dimensions with PCA. Samples generated by our SCI method are more scattered than GDFQ samples and close to that of real data. The additional effect of SCI that diversifies samples by inhibiting feature correlation. We also visualize the synthetic samples of our method and other generative data-free quantization methods (ZeroQ, GDFQ) in Fig. 9 to visually show the effect of data diversifying.
In Fig. 7, it can be obviously investigated that DSQ samples behave more like real data than vanilla data on the offset of mean and standard deviation statistics, which corresponds to the diversity at the distribution level of our generated data scheme. Especially, the SDA plays an important role to make the distribution diverse by slacking the constraint of statistics during the generation process. This phenomenon proves that our DSG scheme diversifies the synthetic data at the distribution level. Moreover, the DSG scheme also generates data with a larger variance compared to the vanilla scheme, which implies that our data samples are widely dispersed and more in line with the real situation. Especially, both SDA and LSE jointly promote diversity at the sample level, which might be useful in providing more content information.

The motivation of the SCI method is to inhibit the correlation among features in the generator, and thus the synthetic samples are separated, as shown in Fig. 8. We visualize the data generated by the vanilla method (GDFQ) and our SCI method, as well as the real data and Gaussian random data. At the right-bottom corner of each subplot, we measure the diversity of each set of features by summarizing the elements in the similarity kernel \( k_f \) as an index \( s = \sum_{i,j \in e_k} [k_{i,j}] \), which indicates the feature distance between samples in spatial perspective. The larger the \( s \), the more similar the samples. As can be seen from Fig. 8, the Gaussian random data is dispersive and stochastic. And it is also quite obvious that the real data is naturally distinct. However, samples generated by the vanilla GDFQ method, as shown in the subplot (a), are generally more concentrated in comparison, some of which are even overlapped. So it has the biggest \( s \). Instead, our SCI utilizes the random Gaussian data as initialization and inhibits the correlation between samples to make the data more dispersed than Gaussian sampling. The samples generated by our method, which has the smallest \( s \), are more dispersed than those generated by the vanilla method and seem close to real data. It demonstrates that the SCI method helps to avoid homogenization from the spatial perspective, and thus diversifies the synthetic samples from each other.

We visualize some synthetic samples of our method and other generative data-free quantization methods (ZeroQ, GDFQ). Take a closer look at Fig. 9, comparing pictures of DSG-PTQ and ZeroQ above, both of which generate images directly and update them iteratively. It is obvious that pictures generated by our DSG-PTQ method seem to have diverse colors with fine-grained textures and coarse-grained figures, while pictures of ZeroQ seem identical and have little difference in between. Meanwhile, below two sets of pictures are generated by DSG-QAT and GDFQ respectively. Owing to the generator, there are much more possibilities in the generation process and finally exhibit in both sets of images. It can be seen with the naked eye that samples of our DSG-QAT have distinct colors with higher saturation degrees, and white images inside the pictures have uncertain patterns. However, although samples generated by GDFQ have different colors and details inside the pictures as well, these samples have a uniform style in texture, which is tanglesome and poor in diversity.

2) Evaluation With Calibration Methods: Additionally, to verify the versatility and robustness of the performance gains of our DSG method, we evaluate our DSG scheme with different calibration methods and compare it with other data generation methods under the same setting. The calibration methods include Percentile, EMA, and MSE. As shown in Table VI, no matter which calibration method is integrated, our DSG scheme substantially outstrips ZeroQ by 6.52%, 9.36%, and 0.61%, respectively. The results strongly suggest that DSG can achieve the leading performance in various experimental settings, and thus the improvement is versatile and robust for various calibration methods.

We also evaluate the synthetic data with DFQ [43], which is a calibration method for data-free quantization. Specifically, DFQ has proposed cross-layer range equalization to equalize the different channel ranges of weight in per-layer quantization and bias correction which is to eliminate the biased quantization error. Both of the two techniques rely on the statistics of BN layers following the convolution layer. Therefore, BN layers are needed to calibrate the corresponding activations, so they have to proceed behind each convolution layer, which results in DFQ only working on specific network architectures and cannot be commonly practiced. Fortunately, generative methods, such as ZeroQ and our DSG, can work on arbitrary architectures, and the statistics of activations can take the place of BN statistics and thus be used in the DFQ method. Table VII shows the closeups of two generative data-free quantization methods, i.e., the ZeroQ methods.
and DSG, in conjunction with the DFQ method. Results show that our DSG outperforms ZeroQ by 0.57% and 3.13% in W6A6 and W8A8 cases.

3) Evaluation With Data-Driven Quantization Methods: The experiments above are conducted on calibration methods that optimize the clipping value for activations. We further evaluate our DSG data with a novel data-driven PTQ method named Adaround [42], which utilizes a rounding approach to quantize weights. Meanwhile, we introduce two other data generation tricks into our DSG scheme, e.g., generating data with labels [25] and DSG, in conjunction with the DFQ method. Results show that our DSG outperforms ZeroQ by 0.57% and 3.13% in W6A6 and W8A8 cases.

V. CONCLUSION

In this paper, we first revisit the sample generation process in generative data-free PTQ and QAT quantization and then give a theoretical analysis that the diversity of synthetic samples is crucial for the data-free quantization and reveal the homogenization of synthetic data in the distribution and sample levels. In this paper, we propose a novel Diverse Sample Generation (DSG) scheme for generative data-free quantization, to address the deficiencies of previous methods which severely debases the quality of the synthetic data and further harms the performance of the quantized network. Our scheme has been evaluated on a variety of bit widths and neural architectures, and the results forcefully demonstrate the effectiveness and versatility of the DSG scheme. It shows notable accuracy improvements in ultra-low bit-width cases (e.g., W4A4). Moreover, benefiting from the enhanced diversity, the performance of the network is significantly improved in various methods integrating with synthetic data, which demonstrates that diversity is an important property of high-quality synthetic data. We hope our work can provide directions for future research on data-free quantization.

REFERENCES

[1] R. Banner, Y. Nahshan, E. Hoffer, and D. Soudry, "Post-training 4-bit quantization of convolution networks for rapid-deployment," 2018, arXiv: 1810.05723.
[2] R. Banner, Y. Nahshan, E. Hoffer, and D. Soudry, "Post-training 4-bit quantization of convolution networks for rapid-deployment," in Proc. 33rd Int. Conf. Neural Inf. Process. Syst., 2019, pp. 7950–7958.
[3] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "MixMatch: A holistic approach to semi-supervised learning," in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 5049–5059.
[4] A. Borodin, “Determinantal point processes,” 2009, arXiv:0911.1153.
[5] A. Bulat and G. Tzimiropoulos, “Hierarchical binary CNNS for landmark localization with limited resources,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 2, pp. 343–356, Feb. 2020.
[6] Y. Cai, Z. Yao, Z. Dong, A. Gholami, M. W. Mahoney, and K. Keutzer, “ZeroQ: A novel zero shot quantization framework,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 13169–13178.
QIN et al.: DIVERSE SAMPLE GENERATION: PUSHING THE LIMIT OF GENERATIVE DATA-FREE QUANTIZATION

[33] C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos, “MMD GAN: A novel approach to MMD estimation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 3552–3556.

[34] A. Mordvintsev, C. Olah, and M. Tyka, “Inceptionism: Going deeper into networks,” in Proc. Google Res. Blog, 2015.

[35] H. Qin, R. Gong, X. Liu, X. Bai, J. Song, and N. Sebe, “Binary neural networks: A survey,” Pattern Recognit., vol. 88, pp. 608–624, 2018.

[36] H. Qin et al., “BiPointNet: Binary neural network for point clouds,” 2020, arXiv:2010.05501.

[37] H. Qin, R. Gong, X. Liu, X. Bai, J. Song, and N. Sebe, “Binary neural networks: A survey,” Pattern Recognit., vol. 105, 2020, Art. no. 107281.

[38] H. Qin et al., “Forward and backward information retention for accurate binary neural networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 2247–2256.

[39] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. 28th Int. Conf. Neural Inf. Process. Syst., 2016, pp. 91–99.

[40] M. Sandler, A. G. Howard, M. Z. A. Zhimoing, and L.-C. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 4510–4520.

[41] A. Shoukai et al., “Generative low-bitwidth data free quantization,” in Proc. Eur. Conf. Comput. Vis., 2020.

[42] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556.

[43] W. Sung, S. Shin, and K. Hwang, “Resiliency of deep neural networks under quantization,” 2015, arXiv:1511.06489.

[44] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 2818–2826.

[45] Z. Tang, X. Peng, K. Li, and D. N. Metaxas, “Towards efficient U-nets: A coupled and quantized approach,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 3, pp. 568–579, Mar. 2020.

[46] J. E. Van Engelen and H. H. Hoos, “A survey on semi-supervised learning,” Mach. Learn., vol. 109, no. 2, pp. 373–440, 2020.

[47] W. Wang, J. Lu, Z. Wu, and J. Zhou, “Learning efficient binarized object detectors with information compression,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 6, pp. 3082–3095, Jun. 2022.

[48] Z. Wang, J. Lu, and J. Zhou, “Learning channel-wise interactions for binary convolutional neural networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 10, pp. 3432–3445, Sep. 2022.

[49] Wikipedia Contributors, “List of ones — Wikipedia, the free Encyclopedia,” 2023, Accessed: May 7, 2023. [Online]. Available: https://en.wikipedia.org/w/index.php
[62] H. Yin et al., “Dreaming to distill: Data-free knowledge transfer via deepinversion,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8715–8724.

[63] X. Zhang et al., “Diversifying sample generation for accurate data-free quantization,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15653–15662.

[64] X. Zhang, X. Zhou, M. Lin, and J. Sun, “ShuffleNet: An extremely efficient convolutional neural network for mobile devices,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6848–6856.

[65] R. Zhao, Y. Hu, J. Dotzel, C. De Sa, and Z. Zhang, “Improving neural network quantization without retraining using outlier channel splitting,” 2019, arXiv: 1901.09504.

[66] B. Zhuang, C. Shen, M. Tan, L. Liu, and I. Reid, “Structured binary neural networks for accurate image classification and semantic segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 413–422.

[67] B. Zhuang, M. Tan, J. Liu, L. Liu, I. Reid, and C. Shen, “Effective training of convolutional neural networks with low-bitwidth weights and activations,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 10, pp. 6140–6152, Oct. 2022.

Haotong Qin received the BEng degree in computer science and engineering from Beihang University. He is currently working toward the PhD degree with the State Key Laboratory of Software Development Environment (SKLSDE) and Shen Yuan Honors College, Beihang University, supervised by Prof. Wei Li and Prof. Xianglong Liu. He is interested in hardware-friendly deep learning and neural network quantization. And his research goal is to enable state-of-the-art neural network models to be deployed on resource-limited hardware, and acceleration for multiple architectures, and the flexible, and efficient deployment on multiple hardware.

Yifu Ding received the BEng degree in computer science and engineering from Beihang University. She is currently working toward the PhD degree with the State Key Laboratory of Software Development Environment (SKLSDE) and Shen Yuan Honors College, Beihang University, under the supervision of Prof. Xianglong Liu. Her research interests include model quantization and acceleration, and getting large state-of-the-art models to perform well on resource-limited edge scenarios.

Xiangguo Zhang received the bachelor’s degree from Shandong University, in 2019, and the master’s degree from the School of Computer Science of Beihang University, under the guidance of Prof. Xianglong Liu. Currently, he works with the Model Toolchain Team for SenseTime Research. He is interested in computer vision and post-training quantization.

Jiakai Wang received the BSc degree from Beihang University, in 2018, and the PhD degree from Beihang University, in 2022, supervised by Prof. Wei Li and Prof. Xianglong Liu. He is now a research scientist with Zhongguancun Laboratory, Beijing, P.R.China. His research interests include trustworthy AI, which consists of physical adversarial examples generation, and adversarial defense and evaluation.

Jiakai Wang received the BEng degree in computer science and engineering from Beihang University. She is currently working toward the PhD degree with the State Key Laboratory of Software Development Environment (SKLSDE) and Shen Yuan Honors College, Beihang University, under the supervision of Prof. Xianglong Liu. Her research interests include model quantization and acceleration, and getting large state-of-the-art models to perform well on resource-limited edge scenarios.

Xianglong Liu (Member, IEEE) received the BS and PhD degrees under the supervision of Prof. Wei Li and visited DVMM Lab, Columbia University as a joint PhD student supervised by Prof. Shih-Fu Chang. He is a full professor with the School of Computer Science and Engineering, Beihang University. His research interests include fast visual computing (e.g., large-scale search/understanding) and robust deep learning (e.g., network quantization, adversarial attack/defense, and few-shot learning).

Jiwen Lu (Senior Member, IEEE) received the PhD degree in electrical and electronic engineering from Nanyang Technological University, Singapore. He is an Associate Professor with the Department of Automation, Tsinghua University, China. His research interests include computer vision (object detection, scene understanding, visual forensics), machine learning (foundation models, representation learning, similarity learning), and unmanned systems (autonomous driving, robotic grasping, visual navigation).