Zero-shot Quantization: A Comprehensive Survey

Minjun Kim[1] , **Jaehyeon Choi**[1] , **Jongkeun Lee**[2] , **Wonjin Cho**[2] and **U Kang**[1,2]
[1]Department of Computer Science and Engineering, Seoul National University
[2]Interdisciplinary Program in Artificial Intelligence, Seoul National University
{minjun.kim, jaehyeon_choi, jklee2, chowonjin0627, ukang}@snu.ac.kr

Abstract

Network quantization has proven to be a powerful approach to reduce the memory and computational demands of deep learning models for deployment on resource-constrained devices. However, traditional quantization methods often rely on access to training data, which is impractical in many real-world scenarios due to privacy, security, or regulatory constraints. Zero-shot Quantization (ZSQ) emerges as a promising solution, achieving quantization without requiring any real data. In this paper, we provide a comprehensive overview of ZSQ methods and their recent advancements. First, we provide a formal definition of the ZSQ problem and highlight the key challenges. Then, we categorize the existing ZSQ methods into classes based on data generation strategies, and analyze their motivations, core ideas, and key takeaways. Lastly, we suggest future research directions to address the remaining limitations and advance the field of ZSQ. To the best of our knowledge, this paper is the first in-depth survey on ZSQ.

1 Introduction

How can we accurately quantize a pre-trained model without any data? Recent advancements in deep neural networks, including architectures like Convolutional Neural Networks (CNNs) [He *et al.*, 2016] and Vision Transformers (ViTs) [Dosovitskiy *et al.*, 2020], have achieved the state-of-the-art results in various applications ranging from image classification to visual question answering [Liu *et al.*, 2023]. However, deploying these models on resource-constrained edge devices remains a significant challenge due to their high memory and computational requirements. Model compression has emerged as a key technique to address these challenges, offering solutions that reduce model size and computational demands [Gholami *et al.*, 2022; Jang *et al.*, 2023; He and Xiao, 2024]. Among various compression strategies, network quantization [Li *et al.*, 2021a; Piao *et al.*, 2022] stands out by converting high-precision models into a low-bit format, offering high compression and faster inference with minimal performance drop, compared to alternatives such as pruning [Park *et al.*, 2024a; He and Xiao, 2024; Park *et al.*, 2025] and knowledge distillation [Tran *et al.*, 2022; Cho and Kang, 2022; Xie *et al.*, 2023; Jeon *et al.*, 2023a].

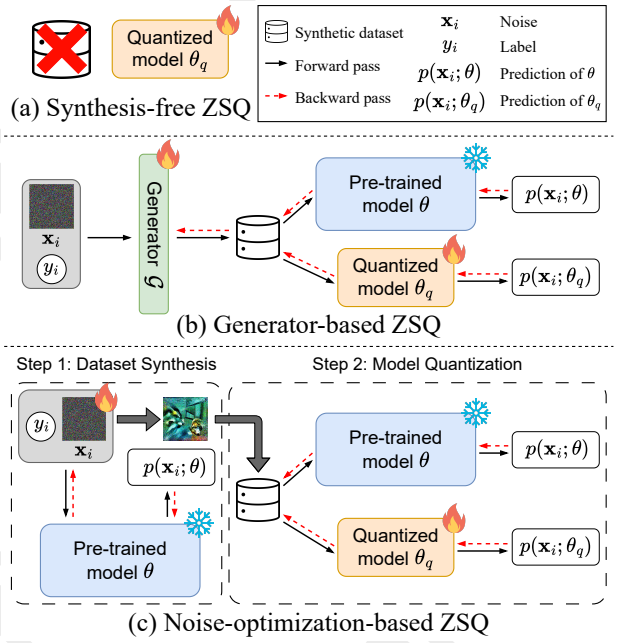

Figure 1: Comparison between three categories of Zero-shot Quantization (ZSQ): (a) synthesis-free, (b) generator-based, and (c) noise-optimization-based ZSQ methods.

Zero-shot Quantization (ZSQ) [Nagel *et al.*, 2019], also called data-free quantization, addresses a critical limitation in traditional quantization techniques: the dependence on training data. This is particularly valuable in scenarios where access to original training datasets is restricted due to privacy, security, or regulatory concerns [Sharma *et al.*, 2021]. These limitations are especially pronounced in industries like healthcare, finance, and enterprise services, where sensitive or proprietary data cannot be leveraged for additional model calibration. Moreover, ZSQ enables the deployment of personalized AI systems, such as mobile AI assistants or security cameras, ensuring efficient performance without compromising data privacy or requiring access to sensitive datasets.

Building upon foundational works [Nagel *et al.*, 2019; Yoo *et al.*, 2019; Xu *et al.*, 2020], numerous studies have further developed the field of ZSQ. With the rapid growth in

research, it has become challenging for researchers to understand the overall trends and essential takeaways of individual studies. However, existing surveys focus on broader topics such as general model compression [Deng *et al.*, 2020; Park *et al.*, 2024b] or network quantization [Gholami *et al.*, 2022; Li *et al.*, 2024a], offering only a brief exploration of ZSQ as a subtopic. This limited coverage restricts researchers from exploring current extent of ZSQ research, distinguishing their key findings, and determining directions for future research.

In this paper, we conduct a comprehensive and structured survey of ZSQ methods. We start by formulating the ZSQ problem and exploring three critical challenges. Next, we summarize existing ZSQ methods (see Table 1) and categorize them based on their data generation strategies: synthesis-free, generator-based, and noise-optimization-based methods. We illustrate the overall process of each category in Figure 1. Then, an in-depth examination of these approaches follows, highlighting their motivations, main ideas, and key insights. Lastly, we propose promising directions for future research, emphasizing unexplored challenges and application scenarios. Our contributions are summarized as follows:

- **Overview.** We identify major trends in ZSQ algorithms, covering diverse data generation approaches and training scenarios (see Figure 1 and Table 1).
- **Analysis.** We provide a comprehensive review of current ZSQ algorithms, highlighting their motivations, principal ideas, and key findings (see Sections 4, 5, and 6).
- **Discussion.** We outline future research directions to advance ZSQ, aiming to guide research toward impactful advancements over current limitations (see Section 7).

2 Problem Formulation

We describe the preliminaries, formulate the ZSQ problem, and discuss the three major challenges that arise in solving it.

2.1 Preliminaries

Network Quantization. Network quantization improves the memory usage and computational efficiency of neural networks by encoding the weight and activations of a given higher-bit network within a lower-bit format. Quantizing a matrix $\mathbf{W}$ to Bbit precision involves first rescaling its values to fit within the interval $[-2^{B-1}, 2^{B-1}-1]$. Each weight is then discretized by mapping to the nearest available integer [Gupta *et al.*, 2015]. Given a matrix $\mathbf{W}$, the Bbit quantized matrix $\mathbf{W}_q$ is calculated as shown in Equation 1.

$$\mathbf{W}_q = \lfloor \frac{\mathbf{W}}{s} + z + \frac{1}{2} \rfloor, \tag{1}$$

where scaling factor $s = (\beta - \alpha)/(2^B - 1)$, zero-point $z = -2^{B-1} - \alpha/s$, and $[\alpha, \beta]$ denotes the clipping range. Properly choosing the clipping range $[\alpha, \beta]$ is essential as it defines s and z required for accurate quantization. A straightforward yet widely adopted approach, known as *min-max quantization*, involves setting α and β to the minimum and maximum values of $\mathbf{W}$, respectively.

QAT and PTQ. Quantization methods are divided into Quantization-aware Training (QAT) and Post-training Quantization (PTQ) based on their training requirements. QAT incorporates quantization during fine-tuning, optimizing model performance under quantization constraints while demanding greater computational resources. In contrast, PTQ is a lightweight approach that quantizes a pre-trained model without additional fine-tuning, making it fast but prone to performance drop. In ZSQ, while QAT methods rely on min-max quantization, PTQ methods employ advanced techniques such as adaptive rounding [Nagel *et al.*, 2020], block reconstruction [Li *et al.*, 2021a], and random dropping [Wei *et al.*, 2022], often leading to better performance.

2.2 Zero-shot Quantization

Given a pre-trained model, Zero-shot Quantization (ZSQ) problem aims to perform network quantization without relying on any real data. The pre-trained model may address different target tasks, such as image classification, object detection, and semantic segmentation. We provide the formal definition in Problem 1.

Problem 1 (Zero-shot Quantization)**.** *We have a model θ trained on a task $\mathcal{T}$ and quantization bits B. The goal is to generate a quantized model θ_q within a Bbit limit without the use of real data, which shows the best performance on $\mathcal{T}$.*

2.3 Main Challenges of ZSQ

Addressing ZSQ requires overcoming key challenges that arise due to the absence of real data. This survey highlights how existing approaches tackle the following challenges:

- **Knowledge transfer from the pre-trained model.** ZSQ relies solely on the information contained in the pre-trained model to recover quantization errors in the absence of real data. Therefore, effectively transferring the knowledge, features, or intrinsic characteristics of the pre-trained model to the quantized model is necessary. Ensuring this transfer without performance degradation remains a critical issue.
- **Discrepancy between real and synthetic datasets.** With no access to real datasets, most ZSQ approaches generate synthetic datasets to fine-tune or calibrate the quantized model. However, synthetic datasets differ significantly from real-world datasets in various aspects, resulting in severe performance degradation. Thus, reducing these disparities to improve the quality of synthetic datasets is crucial.
- **Diversity of the problem setting.** The scope of the ZSQ problem covers a wide range of tasks, model architectures, experimental conditions, and quantization schemes. Hence, it is essential yet challenging to design ZSQ methods that are not only tailored to particular scenarios but also adaptable to diverse tasks or domains.

3 Categorization

We categorize existing ZSQ algorithms based on the data generation approach of each algorithm as synthesis-free (see Section 4), generator-based (see Section 5), and noise-optimization-based (see Section 6) methods. Furthermore, for each category, the methodologies are divided into zero-shot QAT and zero-shot PTQ based on their training requirements.

We summarize the key features of ZSQ methods and compare their performance in Table 1. “Scope of Contribution”

Method	Venue	Training Requirement	Scope of Contribution	Architecture	# Images	Accuracy (FP = 71.47)	
						W4A4	W3A3
Synthesis-free ZSQ							
DFQ [2019]	ICCV	PTQ	Q	CNN	0	55.78	-
SQuant [2022]	ICLR	PTQ	Q	CNN	0	66.14	25.74
UDFC [2023]	ICCV	PTQ	Q	CNN	0	63.49	-
Generator-based ZSQ							
GDFQ [2020]	ECCV	QAT	S, Q	CNN	1.28M	60.60	20.23
ZAQ [2021]	CVPR	QAT	S, Q	CNN	1.28M	52.64	-
ARC [2021]	IJCAI	QAT	S, Q	CNN	1.28M	61.32	23.37
Qimera [2021]	NeurIPS	QAT	S, Q	CNN	1.28M	63.84	1.17
ARC + AIT [2022]	CVPR	QAT	Q	CNN	1.28M	65.73	-
AdaSG [2023b]	AAAI	QAT	S, Q	CNN	1.28M	66.50	37.04
AdaDFQ [2023a]	CVPR	QAT	S, Q	CNN	1.28M	66.53	38.10
Causal-DFQ [2023]	ICCV	QAT	S, Q	CNN	1.28M	68.11	-
RIS [2024]	AAAI	QAT	S	CNN	1.28M	67.75	-
GenQ [2024b]	ECCV	PTQ / QAT	S	CNN	1K§	69.77§	-
Noise-optimization-based ZSQ							
DeepInversion [2020]	CVPR	QAT	S	CNN	32	70.27*	64.28†
IntraQ [2022]	CVPR	QAT	S, Q	CNN	5.12K	66.47	45.51
HAST [2023]	CVPR	QAT	S, Q	CNN	5.12K	66.91	51.15
TexQ [2023]	NeurIPS	QAT	S, Q	CNN	5.12K	67.73	50.28
PLF [2024]	CVPR	QAT	Q	CNN	5.12K	67.02	-
SynQ [2025b]	ICLR	QAT	Q	CNN / ViT	5.12K	67.90	52.02
ZeroQ [2020]	CVPR	PTQ	S, Q	CNN	1K	26.04	-
KW [2020]	CVPR	PTQ	S, Q	CNN	1K	69.08	-
DSG [2021]	CVPR	PTQ	S	CNN	1K	34.53	-
MixMix [2021b]	ICCV	PTQ / QAT	S	CNN	1K§	69.46§	-
PSAQ-ViT [2022]	ECCV	PTQ	S	ViT	32	71.56*	65.57†
Genie [2023b]	CVPR	PTQ	S, Q	CNN	1K	69.66	66.89
SADAG [2024]	ICML	PTQ	S, Q	CNN	1K	69.72	67.10
SMI [2024]	ICML	PTQ	S	ViT	32	70.13*	64.04†
CLAMP-ViT [2024]	ECCV	PTQ	S, Q	ViT	32	72.17*	69.93†

∗ W8A8 accuracy of DeiT-Tiny [2021] model, † W4A8 accuracy of DeiT-Tiny [2021] model, § PTQ setting.

Table 1: A summary of Zero-shot Quantization (ZSQ) methods. W*B*A*B* indicates that weights and activations are quantized to *B*bit. We compare the ZSQ accuracy [%] of a ResNet-18 model pre-trained on ImageNet dataset. See Section 3 for details.

identifies whether the main contribution of each method is for data synthesis (S) or network quantization (Q). "Architecture" specifies the type of neural network architecture the method is applied to, such as CNN or ViT. Also, we evaluate the 3- and 4-bit ZSQ accuracy of a ResNet-18 model [He *et al.*, 2016] pre-trained on ImageNet dataset [Deng *et al.*, 2009] for fair benchmarking across methods. We provide the total number of synthetic samples required to achieve the reported performance as "# Images". For ViT-specific approaches, their ZSQ performance of a DeiT-Tiny model [Touvron *et al.*, 2021] pre-trained on the ImageNet dataset is reported.

4 Synthesis-free ZSQ

Synthesis-free ZSQ methods compress a pre-trained model without generating any synthetic data. These approaches mitigate quantization-induced performance degradation by leveraging structural properties and theoretical foundations of the pre-trained model without generating any synthetic data.

DFQ [Nagel *et al.*, 2019] is a per-tensor weight quantization method that minimizes quantization error through cross-layer equalization and bias correction. Per-tensor quantization inherently results in higher quantization errors for channels with narrow value ranges when grouped with broader-range channels, as distinct weights are compressed into overly coarse bins. DFQ addresses this challenge by equalizing the ranges of channel pairs in consecutive layers and adjusting the bias term of each layer based on batch normalization statistics. DFQ is the first approach to perform weight quantization without any datasets.

SQuant [Guo *et al.*, 2022] enhances the computational efficiency of Hessian-based quantization by exploiting the structural characteristics of CNNs. Hessian-based methods effectively optimize quantized models by evaluating the quantization error with the second-order derivative matrix of each weight, but they suffer from substantial computational overhead due to extensive matrix operations. SQuant improves computational efficiency by performing diagonal Hessian approximation at multiple levels and reduces quantization error by selectively flipping weight signs in decomposed Hessian matrices. SQuant critically improves the efficiency

of Hessian-based quantization without data dependency.

UDFC [Bai *et al.*, 2023] proposes a hybrid model compression algorithm that integrates pruning and quantization techniques. UDFC addresses the damage resulting from pruning or quantization by leveraging the weighted combination of remaining undamaged channels. Specifically, after pruning or quantizing the lth layer, UDFC adjusts the $(l+1)$th layer through a theoretically derived closed-form solution to reduce reconstruction error. UDFC introduces the first unified approach for combining both quantization and pruning in a zero-shot compression setting.

5 Generator-based ZSQ

Generator-based ZSQ methods employ an independent generator model $\mathcal{G}$ to produce synthetic datasets for quantizing pre-trained models. Specifically, they generally train a Generative Adversarial Network (GAN)-based generator [Goodfellow *et al.*, 2014] such as DCGAN and ACGAN from scratch. An ideal generator $\mathcal{G}$ would generate a synthetic dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ with a distribution that closely resembles a real dataset. Therefore, generator-based algorithms aim to train a generator $\mathcal{G}$ to produce synthetic datasets that closely match real data distributions by incorporating crucial information extracted from pre-trained models. One example of such information that is commonly important for quantizing CNN models is Batch Normalization Statistics (BNS). BNS refers to the mean and variance calculated during CNN training, which reflects the distribution of the training data processed by each batch normalization layer. Many studies reduce the L2 norm between the BNS of real and synthetic datasets across all L layers, enforcing a regularization constraint on the synthetic dataset during updates.

5.1 Generator-based Zero-shot QAT

Most generator-based algorithms generate images at each step to refine a generator model to mimic real data accurately. Instead of fine-tuning the quantized model only with the final images from the trained generator, QAT methods exploit images produced at every training stage of the generator as they inherently retain the pre-trained model's knowledge. Specifically, these approaches involve alternating or adversarial learning to balance quantization efficiency and output quality. Generator-based QAT methods share an identical experimental setup with 1.28M synthetic images produced over 400 epochs, each consisting of 200 batches of 16 samples.

GDFQ [Xu *et al.*, 2020] is the first ZSQ method that incorporates a generator to produce a synthetic dataset. GDFQ refines the generator and the quantized model through alternating optimization. The generator aims to produce synthetic data that allows the pre-trained model to predict labels accurately, while the quantized model learns to classify these synthetic images correctly to minimize the performance gap. This strategy progressively improves both data quality and model performance, achieving robust performance.

ZAQ [Liu *et al.*, 2021] employs adversarial learning to optimize quantization by introducing competition between a generator and a quantized model. Adversarial learning follows a minimax optimization, where the generator maximizes the performance gap between the pre-trained and quantized models while the quantized model refines itself to minimize it. Additionally, ZAQ incorporates activation regularization to guide the generator and feature-level knowledge distillation to better encourage the quantized model to mimic the pre-trained model. This adversarial learning framework serves as a basis for generator-based ZSQ research, accelerating the advancement of optimization algorithms in this domain.

ARC [Zhu *et al.*, 2021] or AutoReCon automatically determines a generator architecture by leveraging neural architecture search. Most existing techniques employ GAN-based architectures tailored for image generation instead of model compression. To resolve this issue, ARC employs neural architecture search to find generator architectures suited for compression. With an optimal generator, ARC outperforms existing methods in 3-bit quantization by up to 11%p.

Qimera [Choi *et al.*, 2021] trains a generator to produce boundary supporting samples, aiming to improve quantization performance. Boundary supporting samples are located near the decision boundaries of the pre-trained model; however, existing methods lack such samples in synthetic datasets, which limits the ability of the quantized model to learn the decision boundaries of the pre-trained model effectively. To address this limitation, Qimera encourages the generator to synthesize samples near the decision boundaries by utilizing superposed latent embeddings. Qimera, when combined with existing ZSQ methods, improves up to 9%p in quantization performance under 4-bit quantization settings.

AIT [Choi *et al.*, 2022] emphasizes that Cross-Entropy (CE) loss hinders the optimization process when training quantized models with synthetic datasets. While previous approaches employ both CE and Kullback–Leibler divergence (KL) losses to optimize quantized models, the authors observe two key points: 1) the conflict between CE and KL losses, and 2) the superior generalizability of KL loss. Motivated by these observations, AIT eliminates the CE loss and focuses exclusively on KL loss to optimize the quantized model. Additionally, it manipulates gradients to ensure that a minimum ratio of integer values in the quantized model is updated during each optimization step, thereby enhancing optimization efficiency. AIT integrates easily with other generator-based ZSQ methods to improve performance and promote optimization efficiency.

AdaSG [Qian *et al.*, 2023b] measures 'sample adaptability,' which is the contribution of synthetic images in training quantized models, and proposes a novel optimization algorithm for this metric. Existing ZSQ methods fail to fully recover performance degradation caused by quantization since they neglect the characteristics of quantized models during synthetic image generation. AdaSG incorporates sample adaptability by reformulating the ZSQ problem into a zero-sum game between the generator and the quantized model. AdaSG presents the first game-theoretical formulation of the ZSQ problem, establishing a novel problem-solving paradigm based on sample adaptability.

AdaDFQ [Qian *et al.*, 2023a] proposes a boundary-based optimization method that achieves stable optimization through effectively controlling sample adaptability. Existing ZSQ methods that incorporate sample adaptability encounter

optimization instability, leading to either overfitting or underfitting issues. AdaDFQ defines two boundaries determined by agreement and disagreement based on the predictions of pre-trained and quantized models, and optimizes the margin between these boundaries to ensure that generated samples maintain adaptability with respect to the quantized model. AdaDFQ enhances the stability of the optimization process in ZSQ while maintaining sample adaptability through its boundary-based optimization approach.

Causal-DFQ [Shang *et al.*, 2023] introduces causal reasoning to disentangle content and style for improving synthetic dataset quality. Content captures task-relevant features, while style represents irrelevant attributes that do not influence model decisions. Unlike existing methods that rely on only statistical information (e.g., BNS), Causal-DFQ models these factors separately by constructing a causal graph. It designs a content-style-decoupled generator to synthesize images by independently modulating content and style. This is the first approach to introduce causal relationships in ZSQ, enhancing the diversity and robustness of synthetic data by modulating style factors.

RIS [Bai *et al.*, 2024] encourages the generator to produce diverse synthetic images that contain semantic information. The authors observe that synthetic images are more vulnerable to perturbations compared to real images, indicating a lack of semantic information. To inject such information, RIS explicitly models robustness against perturbations at both feature and prediction levels by applying perturbations to synthetic images and weights of pre-trained models. Furthermore, RIS employs soft labels instead of hard labels as the input of the generator in order to facilitate the creation of diverse synthetic images. The paper reveals that robustness to perturbations in generating synthetic images clearly improves the performance of quantized models.

5.2 Generator-based Zero-shot PTQ

Some generator-based ZSQ methods adopt PTQ scheme by generating synthetic datasets using a pre-trained generator, such as diffusion model, as this approach does not require a training process. Consequently, they adopt PTQ scheme effectively, leveraging characteristics of the synthetic datasets and the architecture of the pre-trained model.

GenQ [Li *et al.*, 2024b] introduces a novel approach to synthesizing reliable data using text-to-image diffusion models. Existing methods struggle to generate semantically rich, high-resolution data due to the complexity of mapping low-dimensional labels to high-dimensional images, causing distribution gaps compared to real data. GenQ addresses this challenge by synthesizing a dataset with Stable Diffusion, introducing three filtering techniques to minimize distribution gaps and improve the quality of synthetic data: 1) energy score filtering, identifying in-distribution data by measuring the confidence of model predictions through energy scores, 2) BNS distribution filtering, aligning the activation statistics of synthetic data with those of real data, and 3) patch similarity filtering, ensuring diverse visual representations for ViTs. GenQ pioneers exploiting text-to-image diffusion models for generating synthetic data to tackle ZSQ problem, resulting in high quantization performance across various settings.

6 Noise-optimization-based ZSQ

Noise-optimization-based ZSQ algorithms directly optimize noise to generate the synthetic dataset by iteratively updating the input itself rather than the parameters of a generator model. They universally follow a two-step scheme by first optimizing randomly initialized noise to generate a synthetic dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ with size N (step 1: dataset synthesis) and then quantizing the pre-trained model with those samples (step 2: model quantization). Similar to generator-based methods, BNS loss serves as the baseline loss, aligning the statistics of all L batch normalization layers between synthetic and real datasets. Additionally, Inception Loss (IL) is optimized to reduce the cross-entropy between sample labels $\{y_i\}_{i=1}^N$ and predictions $\{p(\mathbf{x}_i;\theta)\}_{i=1}^N$ of the pre-trained model θ, encouraging the model to incorporate class-specific details. The main focus is to produce a synthetic dataset that mimics the training dataset of the pre-trained model; each algorithm identifies and mitigates the key discrepancies between real and synthetic datasets.

6.1 Noise-optimization-based Zero-shot QAT

Following dataset synthesis, QAT methods first simply quantize the pre-trained model following min-max quantization and then fine-tune it with the synthetic dataset. These methods share a common experimental setting generating a total of 5,120 images, given that a larger synthetic dataset results in higher model accuracy.

DeepInversion [Yin *et al.*, 2020] pioneers the generation of synthetic datasets for knowledge transfer in scenarios without prior data. The method has two objectives: enforcing feature similarity at all layers between real and synthetic datasets with BNS loss, and generating samples that challenge the quantized model but align with the pre-trained model by penalizing output distribution similarities between the models through training competition. DeepInversion is a general framework that generates a synthetic dataset to transfer knowledge from pre-trained models, applicable not only to ZSQ but also to various data-free tasks such as pruning, knowledge distillation, and continual learning.

IntraQ [Zhong *et al.*, 2022] identifies intra-class heterogeneity as a key factor for improving synthetic dataset quality. Most approaches generate consistent samples within each class because they prioritize overall dataset distribution and inter-class separation. IntraQ promotes intra-class diversity by varying object scale and location with local object reinforcement, distributing class features broadly across images with a marginal distance constraint, and mitigating overfitting with soft IL. Their findings stand out through their high performance, advancing beyond traditional methods which face challenges with bit assignments under 4bit, and taking the first step into extreme low-bit ZSQ at 3bit.

HAST [Li *et al.*, 2023] highlights that including hard samples in a synthetic dataset enhances ZSQ performance. Previous methods perform poorly on hard images, where the difficulty of an image [Li *et al.*, 2019] indicates how likely a model is to misclassify it, because their synthetic datasets lack challenging samples. Therefore, HAST increases the portion of difficult samples within the synthetic dataset for

both pre-trained and quantized models by optimizing hard-sample-enhanced IL and promoting sample difficulty, respectively. HAST shows outstanding performance, comparable to that of fine-tuning with real datasets.

TexQ [Chen *et al.*, 2023] focuses on minimizing discrepancies in the distributions of texture features. Texture describes the spatial arrangement of pixel intensities forming visual patterns, vital for CNNs which classify images based on surface characteristics; prior methods often underperform due to insufficient texture features in the synthetic dataset. TexQ resolves this issue by directing samples to accurate per-class calibration centers, ensuring texture alignment with LAWS energy loss, and containing texture features in shallow layers with layered BNS loss. The authors empirically validate that adding sufficient texture features enhances inter-class distance, leading to better ZSQ performance.

PLF [Fan *et al.*, 2024] is the first approach to evaluate the synthetic dataset before quantization. After synthesizing the dataset, PLF separates it into high- and low-reliable groups based on self-entropy computed from the probabilities of a pre-trained model, where lower self-entropy indicates more confident predictions. Then, PLF assigns the second highest probability label as an auxiliary label for low-reliable data to soften supervised learning and reduce the risk from misleading labels. Unlike methods such as HAST, which determine confidence based on image difficulty [Li *et al.*, 2019], PLF adopts self-entropy to measure confidence effectively.

SynQ [Kim *et al.*, 2025b] emphasizes three major limitations when fine-tuning with synthetic datasets: mismatches of amplitude distribution in the frequency domain, predictions based on off-target patterns, and harmful effects of erroneous hard labels for hard samples. SynQ addresses these challenges by introducing a low-pass filter to reduce high-frequency noise, aligning class activation maps to identify correct image regions, and excluding cross-entropy loss for hard samples to reduce ambiguity. SynQ is the first work to explore both CNNs and ViTs, showing high adaptability toward various models and dataset synthesis algorithms.

6.2 Noise-optimization-based Zero-shot PTQ

PTQ algorithms adjust the scaling factor s and zero-point z or apply advanced quantization techniques instead of directly updating model parameters. Compared to QAT, these methods require smaller synthetic datasets for calibration, typically evaluated with sets of 1,000 for CNNs and 32 for ViTs.

ZeroQ [Cai *et al.*, 2020] pioneers zero-shot PTQ as the first method of its kind. ZeroQ first generates a synthetic dataset by optimizing a set of random noise with BNS loss and then employs it to determine the clipping range $[\alpha, \beta]$. This approach further extends to mixed-precision quantization by assigning bit precision configurations based on a Pareto frontier. ZeroQ guides future researches toward enhancing synthetic datasets while maintaining its sample-driven approach for clipping range selection.

KW [Haroush *et al.*, 2020] proposes generating class-specific synthetic datasets by integrating BNS loss and IL. In contrast to previous techniques which do not produce class-specific data, this approach generates images corresponding to each image class. IL aligns the classifier probabilities of images with their pre-assigned labels, thereby injecting desired class information into each image. Combining BNS loss and IL achieves high performance in 4bit quantization, showing only a 2-3%p accuracy drop compared to the original model, making it a baseline loss for future studies.

DSG [Zhang *et al.*, 2021] addresses the homogenization issues in synthetic datasets limited by BNS at distribution and sample levels. The authors report that feature distributions in synthetic datasets overfit to BNS (distribution-level homogenization), and identical optimization objectives result in excessive sample similarity (sample-level homogenization). DSG improves data diversity by slack distribution alignment, which relaxes BNS constraints, and layer-wise sample enhancement, which reinforces per-sample loss. DSG demonstrates that addressing the homogenization caused by BNS significantly reduces overfitting and boosts the diversity of synthetic datasets.

MixMix [Li *et al.*, 2021b] focuses on reducing biases at feature and label levels in generated samples. The authors observe that features are biased as they originate from a specific model, preventing direct application across different architectures. Additionally, labels of generated samples are biased due to inexact inversion from low-dimensional to high-dimensional data. MixMix introduces two mixing techniques: feature mixing to construct a universal feature space across models and data mixing to ensure accurate label representation by combining synthetic samples and labels. The results indicate that synthetic datasets are inherently biased and suitable only for their respective pre-trained models.

PSAQ-ViT [Li *et al.*, 2022] is the first ZSQ method tailored for ViTs. CNN-based methods rely mainly on BNS loss, making them unsuitable for ViTs with layer normalization and self-attention modules. PSAQ-ViT optimizes the Patch Similarity Entropy (PSE) loss, aiming to maximize the diversity of cosine similarities between the outputs of self-attention layers for two different input patches, thereby generating samples that reflect self-attention diversity. PSE loss serves as the baseline loss in ViT-based ZSQ, similar to BNS loss in CNN-based ZSQ.

Genie [Jeon *et al.*, 2023b] integrates generator-based methods with noise-optimization-based algorithms. Genie first trains a generator to extract the common knowledge of the input domain, then indirectly optimizes the synthetic dataset by learning from the trained generator. They distill the knowledge with swing convolution to avoid information loss while maintaining computational efficiency. Genie further adopts advanced PTQ techniques such as adaptive rounding [Nagel *et al.*, 2020], block-wise reconstruction [Li *et al.*, 2021a], and random dropping [Wei *et al.*, 2022], leading to improved performance.

SADAG [Dung *et al.*, 2024] proposes a sharpness-aware generation method to enhance the generalization of the quantized model. Optimizing the noise with sharpness-aware minimization reduces both loss value and sharpness, leading the image to a flat local optimum. Therefore, SADAG creates samples with less loss sharpness by applying gradient matching. The authors find that optimizing for loss sharpness generates smoother images with reduced color variation while maintaining semantics, leading to better performance.

SMI [Hu *et al.*, 2024] generates only the essential parts of an image to accelerate generation speed while maintaining ZSQ performance. Existing methods rely on dense model inversion to produce all pixels of a fixed-size image, unnecessarily spending equal time on unimportant backgrounds and often introducing label errors by generating objects from other classes. SMI addresses this issue by evaluating patch-level importance through ViT attention scores and stopping optimization for irrelevant patches, resulting in a sparse dataset. SMI achieves up to 3.79× faster image generation with 77% sparsity while achieving similar or up to 0.78%p higher accuracy in W4A8 quantization, highlighting the effectiveness of sparse model inversion.

CLAMP-ViT [Ramachandran *et al.*, 2024] adopts contrastive learning to integrate semantic information into the synthetic dataset. The authors point out that patch similarity in PSAQ-ViT assumes equal importance for all patches without considering spatial sensitivity, failing to capture semantically meaningful inter-patch relations. To overcome this problem, CLAMP-ViT introduces a patch-level contrastive learning framework designed to enhance the semantic quality of synthetic data tailored for ViT models. Furthermore, CLAMP-ViT supports mixed-precision quantization through layer-wise evolutionary search, which determines the optimal bit-width and quantization parameters.

7 Further Research Directions

ZSQ is a rapidly growing field with substantial potential. Although notable advances have been made in ZSQ, numerous research directions remain open for future exploration.

- **More principled analysis on synthetic datasets.** As mentioned in Section 2.3, reducing the discrepancy between real and synthetic datasets is one of the significant challenges for ZSQ methods. However, researchers often focus on individual features (e.g., intra-class heterogeneity, amplitude distribution, etc.) instead of investigating the underlying causes of their differences. Conducting a deeper analysis of synthetic datasets might fundamentally enhance ZSQ methods, going beyond mere patchwork solutions.
- **Broader application to various tasks and domains.** While ZSQ encompasses a wide range of settings, most research concentrates on the image domain, focusing primarily on CNN and ViT models. Specifically, they set task $\mathcal{T}$ mainly as image classification, with a few work on object detection [Li *et al.*, 2022; Shang *et al.*, 2023] or semantic segmentation [Nagel *et al.*, 2019]. Extending ZSQ research to various tasks such as language, multi-variate, or graph-based ones is crucial for advancing the field [Kim *et al.*, 2021; Kim *et al.*, 2025a].
- **Theoretical exploration of ZSQ.** A formal investigation into the theoretical limits of zero-shot quantization performance is essential for understanding its full potential. This includes defining the upper bounds on model accuracy without access to real data, and exploring how quantization bit-widths affect performance degradation. Furthermore, identifying the mathematical principles underlying ZSQ is crucial for developing more robust algorithms.
- **Faster generation of synthetic datasets.** Increasing the size of synthetic datasets enhances the performance of quantized models by providing more diverse training data [Kim *et al.*, 2025b]. However, existing algorithms require a significant amount of time, requiring 1 to 4 RTX 4090 GPU hours for generating 5,120 images with a resolution of 224 × 224. Therefore, reducing the time required for data generation is a vital topic for future exploration.
- **Combining other model compression techniques.** Current ZSQ methods achieve competitive results in 4bit quantization but struggle to retain performance in 3bit or lower-bit quantization. Integrating quantization with other model compression methods such as pruning, weight sharing, or low-rank approximation might be a key to achieve higher compression rate while maintaining accuracy.
- **Evaluating practical impact on real-world scenarios.** The importance of ZSQ lies in its applications for handling real-world scenarios with limited data. However, current ZSQ methods present experimental results solely on benchmark datasets and models. The lack of evaluation restricts researchers from validating the practicality of proposed methods, limiting the reliability of ZSQ approaches. Hence, evaluating ZSQ methods under practical settings is required to compare their real-world applicability.
- **Diverse problem settings.** Recent advancements in ZSQ have led to the exploration of diverse problem settings beyond conventional scenarios. These include setups such as few-instance quantization (with 1 to 10 real images) or leveraging pre-trained diffusion models. Extending ZSQ to real-time quantization and edge-device deployments are vital future directions in enhancing the practical utility.

8 Conclusion

In this paper, we perform an extensive survey of ZSQ. We first present the preliminaries and formulate the problem with its key challenges. Then, we review ZSQ algorithms with a novel taxonomy and categorize them based on data generation approaches and training requirements. Specifically, we detail the motivations, ideas, and findings of each paper comprehensively. Finally, we introduce promising research topics on ZSQ, providing insights to resolve current constraints.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [No. RS-2020-II200894, Flexible and Efficient Model Compression Method for Various Applications and Environments], [No. RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)], and [No. RS-2021-II212068, Artificial Intelligence Innovation Hub (Artificial Intelligence Institute, Seoul National University)]. This work was supported by Youlchon Foundation. The Institute of Engineering Research at Seoul National University provided research facilities for this work. The ICT at Seoul National University provides research facilities for this study. Minjun Kim and Jaehyeon Choi contributed equally to this paper. U Kang is the corresponding author.

References

[Bai *et al.*, 2023] Shipeng Bai, Jun Chen, Xintian Shen, Yixuan Qian, and Yong Liu. Unified data-free compression: Pruning and quantization without fine-tuning. In *ICCV*, 2023.

[Bai *et al.*, 2024] Jianhong Bai, Yuchen Yang, Huanpeng Chu, Hualiang Wang, Zuozhu Liu, Ruizhe Chen, Xiaoxuan He, Lianrui Mu, Chengfei Cai, and Haoji Hu. Robustness-guided image synthesis for data-free quantization. In *AAAI*, 2024.

[Cai *et al.*, 2020] Yaohui Cai, Zhewei Yao, Zhen Dong, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Zeroq: A novel zero shot quantization framework. In *CVPR*, 2020.

[Chen *et al.*, 2023] Xinrui Chen, Yizhi Wang, Renao Yan, Yiqing Liu, Tian Guan, and Yonghong He. Texq: zero-shot network quantization with texture feature distribution calibration. In *NeurIPS*, 2023.

[Cho and Kang, 2022] Ikhyun Cho and U Kang. Pea-kd: Parameter-efficient and accurate knowledge distillation on bert. *PLOS ONE*, 17(2):1–12, 02 2022.

[Choi *et al.*, 2021] Kanghyun Choi, Deokki Hong, Noseong Park, Youngsok Kim, and Jinho Lee. Qimera: Data-free quantization with synthetic boundary supporting samples. In *NeurIPS*, 2021.

[Choi *et al.*, 2022] Kanghyun Choi, Hye Yoon Lee, Deokki Hong, Joonsang Yu, Noseong Park, Youngsok Kim, and Jinho Lee. It's all in the teacher: Zero-shot quantization brought closer to the teacher. In *CVPR*, 2022.

[Deng *et al.*, 2009] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[Deng *et al.*, 2020] Lei Deng, Guoqi Li, Song Han, Luping Shi, and Yuan Xie. Model compression and hardware acceleration for neural networks: A comprehensive survey. *Proceedings of the IEEE*, 108(4):485–532, 2020.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020.

[Dung *et al.*, 2024] Hoang Anh Dung, Cuong Pham, Trung Le, Jianfei Cai, and Thanh-Toan Do. Sharpness-aware data generation for zero-shot quantization. In *ICML*, 2024.

[Fan *et al.*, 2024] Chunxiao Fan, Ziqi Wang, Dan Guo, and Meng Wang. Data-free quantization via pseudo-label filtering. In *CVPR*, 2024.

[Gholami *et al.*, 2022] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference. In *Low-Power Computer Vision*, pages 291–326. Chapman and Hall/CRC, 2022.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

[Guo *et al.*, 2022] Cong Guo, Yuxian Qiu, Jingwen Leng, Xiaotian Gao, Chen Zhang, Yunxin Liu, Fan Yang, Yuhao Zhu, and Minyi Guo. Squant: On-the-fly data-free quantization via diagonal hessian approximation. In *ICLR*, 2022.

[Gupta *et al.*, 2015] Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan. Deep learning with limited numerical precision. In *ICML*, 2015.

[Haroush *et al.*, 2020] Matan Haroush, Itay Hubara, Elad Hoffer, and Daniel Soudry. The knowledge within: Methods for data-free model compression. In *CVPR*, 2020.

[He and Xiao, 2024] Yang He and Lingao Xiao. Structured pruning for deep convolutional neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):2900–2919, 2024.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[Hu *et al.*, 2024] Zixuan Hu, Yongxian Wei, Li Shen, Zhenyi Wang, Lei Li, Chun Yuan, and Dacheng Tao. Sparse model inversion: Efficient inversion of vision transformers for data-free applications. In *ICML*, 2024.

[Jang *et al.*, 2023] Jun-Gi Jang, Chun Quan, Hyun Dong Lee, and U Kang. Falcon: lightweight and accurate convolution based on depthwise separable convolution. *Knowl. Inf. Syst.*, 65(5):2225–2249, 2023.

[Jeon *et al.*, 2023a] Hyojin Jeon, Seungcheol Park, Jin-Gee Kim, and U. Kang. Pet: Parameter-efficient knowledge distillation on transformer. *PLOS ONE*, 18(7):1–21, 07 2023.

[Jeon *et al.*, 2023b] Yongkweon Jeon, Chungman Lee, and Ho-young Kim. Genie: show me the data for quantization. In *CVPR*, 2023.

[Kim *et al.*, 2021] Junghun Kim, Jinhong Jung, and U. Kang. Compressing deep graph convolution network with multi-staged knowledge distillation. *PLOS ONE*, 16(8):1–18, 08 2021.

[Kim *et al.*, 2025a] Minjun Kim, Jaehyeon Choi, Seungjoo Lee, Jinhong Jung, and U Kang. Augward: Augmentation-aware representation learning for accurate graph classification. In *PAKDD*, 2025.

[Kim *et al.*, 2025b] Minjun Kim, Jongjin Kim, and U Kang. Synq: Accurate zero-shot quantization by synthesis-aware fine-tuning. In *ICLR*, 2025.

[Li *et al.*, 2019] Buyu Li, Yu Liu, and Xiaogang Wang. Gradient harmonized single-stage detector. In *AAAI*, 2019.

[Li *et al.*, 2021a] Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. Brecq: Pushing the limit of post-training quantization by block reconstruction. In *ICLR*, 2021.

[Li *et al.*, 2021b] Yuhang Li, Feng Zhu, Ruihao Gong, Mingzhu Shen, Xin Dong, Fengwei Yu, Shaoqing Lu, and Shi Gu. Mixmix: All you need for data-free compression are feature and data mixing. In *ICCV*, 2021.

[Li *et al.*, 2022] Zhikai Li, Liping Ma, Mengjuan Chen, Junrui Xiao, and Qingyi Gu. Patch similarity aware data-free quantization for vision transformers. In *ECCV*, 2022.

[Li *et al.*, 2023] Huantong Li, Xiangmiao Wu, Fanbing Lv, Daihai Liao, Thomas H Li, Yonggang Zhang, Bo Han, and Mingkui Tan. Hard sample matters a lot in zero-shot quantization. In *CVPR*, 2023.

[Li *et al.*, 2024a] Min Li, Zihao Huang, Lin Chen, Junxing Ren, Miao Jiang, Fengfa Li, Jitao Fu, and Chenghua Gao. Contemporary advances in neural network quantization: A survey. In *IJCNN*, 2024.

[Li *et al.*, 2024b] Yuhang Li, Youngeun Kim, Donghyun Lee, Souvik Kundu, and Priyadarshini Panda. Genq: Quantization in low data regimes with generative synthetic data. In *ECCV*, 2024.

[Liu *et al.*, 2021] Yuang Liu, Wei Zhang, and Jun Wang. Zero-shot adversarial quantization. In *CVPR*, 2021.

[Liu *et al.*, 2023] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[Nagel *et al.*, 2019] Markus Nagel, Mart van Baalen, Tijmen Blankevoort, and Max Welling. Data-free quantization through weight equalization and bias correction. In *ICCV*, 2019.

[Nagel *et al.*, 2020] Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization. In *ICML*, 2020.

[Park *et al.*, 2024a] Seungcheol Park, Hojun Choi, and U Kang. Accurate retraining-free pruning for pretrained encoder-based language models. In *ICLR*, 2024.

[Park *et al.*, 2024b] Seungcheol Park, Jaehyeon Choi, Sojin Lee, and U Kang. A comprehensive survey of compression algorithms for language models. *arXiv preprint arXiv:2401.15347*, 2024.

[Park *et al.*, 2025] Seungchul Park, Sojin Lee, Jongjin Kim, Jinsik Lee, Hyunsik Jo, and U Kang. Accurate sublayer pruning for large language models by exploiting latency and tunability information. In *IJCAI*, 2025.

[Piao *et al.*, 2022] Tairen Piao, Ikhyun Cho, and U. Kang. Sensimix: Sensitivity-aware 8-bit index & 1-bit value mixed precision quantization for bert compression. *PLOS ONE*, 17(4):1–22, 2022.

[Qian *et al.*, 2023a] Biao Qian, Yang Wang, Richang Hong, and Meng Wang. Adaptive data-free quantization. In *CVPR*, 2023.

[Qian *et al.*, 2023b] Biao Qian, Yang Wang, Richang Hong, and Meng Wang. Rethinking data-free quantization as a zero-sum game. In *AAAI*, 2023.

[Ramachandran *et al.*, 2024] Akshat Ramachandran, Souvik Kundu, and Tushar Krishna. Clamp-vit: contrastive data-free learning for adaptive post-training quantization of vits. In *ECCV*, 2024.

[Shang *et al.*, 2023] Yuzhang Shang, Bingxin Xu, Gaowen Liu, Ramana Rao Kompella, and Yan Yan. Causal-dfq: Causality guided data-free network quantization. In *CVPR*, 2023.

[Sharma *et al.*, 2021] Prasen Kumar Sharma, Arun Abraham, and Vikram Nelvoy Rajendiran. A generalized zero-shot quantization of deep convolutional neural networks via learned weights statistics. *IEEE Transactions on Multimedia*, 25:953–965, 2021.

[Touvron *et al.*, 2021] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021.

[Tran *et al.*, 2022] Cuong Tran, Ferdinando Fioretto, Jung-Eun Kim, and Rakshit Naidu. Pruning has a disparate impact on model accuracy. In *NeurIPS*, 2022.

[Wei *et al.*, 2022] Xiuying Wei, Ruihao Gong, Yuhang Li, Xianglong Liu, and Fengwei Yu. Qdrop: Randomly dropping quantization for extremely low-bit post-training quantization. In *ICLR*, 2022.

[Xie *et al.*, 2023] Yi Xie, Huaidong Zhang, Xuemiao Xu, Jianqing Zhu, and Shengfeng He. Towards a smaller student: Capacity dynamic distillation for efficient image retrieval. In *CVPR*, 2023.

[Xu *et al.*, 2020] Shoukai Xu, Haokun Li, Bohan Zhuang, Jing Liu, Jiezhang Cao, Chuangrun Liang, and Mingkui Tan. Generative low-bitwidth data free quantization. In *ECCV*, 2020.

[Yin *et al.*, 2020] Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *CVPR*, 2020.

[Yoo *et al.*, 2019] Jaemin Yoo, Minyong Cho, Taebum Kim, and U Kang. Knowledge extraction with no observable data. In *Advances in Neural Information Processing Systems*, 2019.

[Zhang *et al.*, 2021] Xiangguo Zhang, Haotong Qin, Yifu Ding, Ruihao Gong, Qinghua Yan, Renshuai Tao, Yuhang Li, Fengwei Yu, and Xianglong Liu. Diversifying sample generation for accurate data-free quantization. In *CVPR*, 2021.

[Zhong *et al.*, 2022] Yunshan Zhong, Mingbao Lin, Gongrui Nan, Jianzhuang Liu, Baochang Zhang, Yonghong Tian, and Rongrong Ji. Intraq: Learning synthetic images with intra-class heterogeneity for zero-shot network quantization. In *CVPR*, 2022.

[Zhu *et al.*, 2021] Baozhou Zhu, Peter Hofstee, Johan Peltenburg, Jinho Lee, and Zaid Alars. Autorecon: Neural architecture search-based reconstruction for data-free compression. In *IJCAI*, 2021.