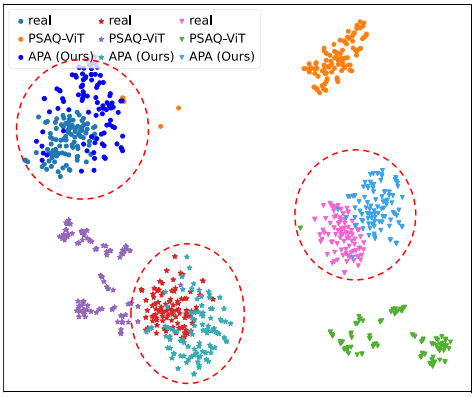
arXiv:2412.16553v5 [cs.CV] 31 Oct 2025

Semantic Alignment and Reinforcement for Data-Free Quantization of Vision Transformers

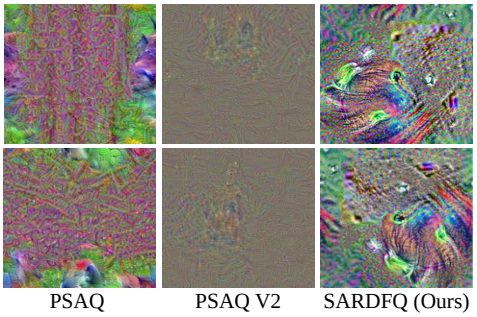
Yunshan Zhong[1,2], Yuyao Zhou[2], Yuxin Zhang[2], Wanchen Sui[3], Shen Li[3],
Yong Li[3],Fei Chao[2], Rongrong Ji[1,2*]
[1]Institute of Artificial Intelligence, Xiamen University
[2]MAC Lab, Department of Artificial Intelligence, School of Informatics, Xiamen University
[3]Alibaba
zhongyunshan@stu.xmu.edu.cn, yuyaozhou@stu.xmu.edu.cn
yuxinzhang@stu.xmu.edu.cn, wanchen.swc@alibaba-inc.com
litan.ls@alibaba-inc.com, jiufeng.ly@alibaba-inc.com, fchao@xmu.edu.cn, rrji@xmu.edu.cn

Abstract

*Data-free quantization (DFQ) enables model quantization without accessing real data, addressing concerns regarding data security and privacy. With the growing adoption of Vision Transformers (ViTs), DFQ for ViTs has garnered significant attention. However, existing DFQ methods exhibit two limitations: (1) semantic distortion, where the semantics of synthetic images deviate substantially from those of real images, and (2) semantic inadequacy, where synthetic images contain extensive regions with limited content and oversimplified textures, leading to suboptimal quantization performance. To address these limitations, we propose SARDFQ, a novel **S**emantics **A**lignment and **R**einforcement **D**ata-**F**ree **Q**uantization method for ViTs. To address semantic distortion, SARDFQ incorporates Attention Priors Alignment (APA), which optimizes synthetic images to follow randomly generated structure attention priors. To mitigate semantic inadequacy, SARDFQ introduces Multi-Semantic Reinforcement (MSR), leveraging localized patch optimization to enhance semantic richness across synthetic images. Furthermore, SARDFQ employs Soft-Label Learning (SL), wherein multiple semantic targets are adapted to facilitate the learning of multi-semantic images augmented by MSR. Extensive experiments demonstrate the effectiveness of SARDFQ, significantly surpassing existing methods. For example, SARDFQ improves top-1 accuracy on ImageNet by 15.52% for W4A4 ViT-B*[1].



(a) The t-SNE[68] visualization of the penultimate-layer features (extracted by DeiT-S) of synthetic images. Each marker (circle, star, triangle) represents a distinct category. The red dashed circles highlight the features extracted from our APA and real images. Notably, the features produced by PSAQ-ViT[47] exhibit substantial deviation from those of real images, indicating semantic distortion. In contrast, our APA yields features that more closely align with those of real images, suggesting improved semantics alignment.



PSAQ PSAQ V2 SARDFQ (Ours)

(b) The images of PSAQ-ViT and PSAQ-ViT V2 exhibit numerous dull regions with limited content and simplified textures, reflecting semantic inadequacy. In comparison, our SARDFQ generates images with greater diversity in both content and texture, demonstrating enhanced semantics.

Figure 1. Illustration of (a) semantic distortion and (b) semantic inadequacy.

*Corresponding Author: rrji@xmu.edu.cn
[1]The code is at https://github.com/zysxmu/SARDFQ.

1. Introduction

Vision Transformers (ViTs) [20, 27, 34] have demonstrated remarkable success across various computer vision tasks [1, 7, 56, 67]. However, their high computational cost and substantial memory footprint hinder deployment in resource-constrained environments [11, 28, 30, 46, 55, 66, 81]. To address these limitations, quantization [39] has emerged as a promising solution, which reduces model complexity by converting full-precision weight and activations into low-bit representations.

Traditional quantization methods typically require access to the original training dataset, which raises data privacy and security concerns [6, 25, 76, 83]. As a result, data-free quantization (DFQ) has gained increasing attention, allowing quantization without the need for real data [14, 75, 80]. However, most existing DFQ methods are designed specifically for convolutional neural networks (CNNs) and are not directly applicable to vision transformers (ViTs). These methods generally rely on batch normalization statistics (BNS), which capture the distribution of real data, to synthesize in-distribution synthetic data [6, 75, 80, 83]. Yet, BNS is unavailable for ViTs, which use layer normalization (LN) to dynamically compute distribution statistics during inference [47]. Recently, several DFQ methods have been proposed for ViTs [16, 31, 47, 48, 62]. For example, PSAQ-ViT [47] introduces patch similarity entropy (PSE) loss to optimize Gaussian noise towards usable synthetic images.

Nevertheless, we observe that existing methods suffer from semantic distortion and inadequacy issues. As shown in Fig. 1a[2], features of synthetic images generated by PSAQ-ViT deviate significantly from those of real images. Tab. 1 shows that the cosine similarity between synthetic images from PSAQ-ViT and real images is notably low, also indicating significant distortion. These results highlight the issue of **semantic distortion**. Moreover, as shown in Fig. 1b[3], synthetic images generated by PSAQ-ViT and PSAQ-ViT V2 exhibit many regions with limited content diversity and overly simplified textures. These low-quality dull regions are useless or even detrimental to model learning [31], highlighting the issue of **semantic inadequacy**. Consequently, quantized models trained on such low-quality images suffer from degraded performance.

Motivated by the above analysis, we propose a novel **S**emantics **A**lignment and **R**einforcement **D**ata-**F**ree **Q**uantization method for ViTs, termed SARDFQ. The overall framework is depicted in Fig. 2. To address the semantic distortion issue, SARDFQ introduces Attention Priors Alignment (APA), where synthetic images are optimized to follow structured attention priors generated using Gaussian Mixture Models (GMMs). APA effectively aligns the semantics of synthetic images with real images, as validated by both visual and quantitative analyses. As shown in Fig. 1a, features of APA exhibit a closer alignment to those of real images, while quantitative results in Tab. 1 also confirm that the semantics of APA are more consistent to the real images, indicating enhanced semantics alignment. To address the semantic inadequacy issue, SARDFQ incorporates Multi-Semantic Reinforcement (MSR) and Softlabel Learning (SL). MSR utilizes localized patch optimization, which encourages different sub-patches of synthetic images to capture various semantics, reinforcing the rich semantics across images. SL applies multiple semantic targets to accommodate the learning of multi-semantic images augmented by MSR. As shown in Fig. 1b, synthetic images after applying MSR exhibit greater diversity in content and texture, providing reinforced semantics.

Experimental results across various ViT models and tasks demonstrate that SARDFQ presents substantial performance improvements. For example, SARDFQ achieves a 15.52% increase in top-1 accuracy on the ImageNet dataset for the W4A4 ViT-B model.

[2]We do not plot the results of images from PSAQ-ViT V2 as they lack specific categories.

[3]More results are presented in the appendix.

2. Related Works

2.1. Vision Transformers

The great success of transformers in the natural language processing field has driven widespread attempts in the computer vision community to apply them to vision tasks [12, 26, 71]. ViT [20] is the pioneer that builds a transformer-based model to handle images, boosting the performance on the image classification task. DeiT [67] introduces an efficient teacher-student training strategy where a distillation token is employed to distill knowledge from the teacher model to the student model. Swin Transformers [54] builds an efficient and effective hierarchical model by introducing a shifted window-based self-attention mechanism. Other than the image classification task, the applications of ViTs also have broadened considerably, manifesting groundbreaking performance in object detection [7], image segmentation [9, 82], low-level vision [50], video recognition [1, 59], and medical image processing [64], *etc*. Nevertheless, the impressive performance of ViTs relies on a high number of parameters and significant computational overhead, preventing deployment in resource-constrained environments. Several recent efforts design lightweight ViTs, such as MobileViT [57], MiniVit [79], and TinyViT [72]. However, the model complexity is still unsatisfactory [47].

2.2. Network Quantization

Data-Driven Quantization. Model quantization reduces the complexity of neural networks by replacing full-precision weight and activation with the low-bit format. Data-driven quantization can be roughly divided into two

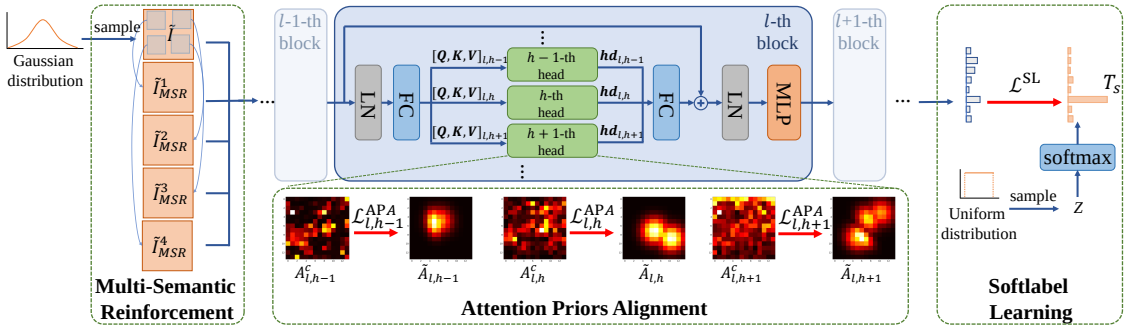


Figure 2. SARDFQ Framework overview: Attention Priors Alignment (APA) employs randomly generated attention priors to improve semantics alignment. Multi-Semantic Reinforcement (MSR) learns the different regions of synthetic images with various semantics to enhance overall semantic richness. Meanwhile, Softlabel Learning (SL) adopts multiple semantic targets to ensure consistent learning of multi-semantic images augmented by MSR.

categories: quantization-aware training (QAT) and post-training quantization (PTQ). QAT is compute-heavy since it re-trains the quantized model with the full training data to retain performance [21, 23, 41, 43, 46, 52, 65, 73]. PTQ perform quantization with a tiny dataset and a reduced time overhead, harvesting widespread attention [3, 42]. The specific architecture of ViTs, such as LayerNorm and the self-attention module, urges distinct PTQ methods compared to CNNs [18, 22, 46, 51, 53, 85]. For example, Liu *et al*. [55] develop a ranking loss to maintain the relative order of the self-attention activation. Unfortunately, both QAT and PTQ involve the original training data, causing concerns about data privacy and security issues in data-sensitive scenarios.

Data-Free Quantization. DFQ quantizes models without accessing real data [2, 13–15, 25, 33, 40, 44, 45, 61]. Most previous DFQ methods focus on CNN, where the BNS can be adopted as the regularization term [6, 80]. However, BNS is infeasible for ViTs built on the LN. Recently, few efforts have been explored to accommodate ViTs [16, 31, 47, 48, 62]. PSAQ-ViT [47] introduces the first DFQ method for ViTs. They discover that Gaussian noise yields homogeneous patches, while the real image yields heterogeneous patches. Thus, patch similarity entropy (PSE) loss is proposed to optimize the Gaussian noise towards real-like images by making them showcase heterogeneous patches. Based on PSAQ-ViT, PSAQ-ViT V2 [48] further introduces an adversarial learning strategy [24]. [62] incorporates contrastive learning and proposes an iterative generation-quantization PTQ-based DFQ method. [31] proposes a sparse generation method to remove noisy and hallucination backgrounds in synthetic images. SynQ [37] introduces an overall framework for DFQ by integrating a low-pass filter, class activation map alignment, and difficult samples learning. For a more comprehensive DFQ survey, we invite the reader to see [36].

3. Method

3.1. Preliminaries

3.1.1. Quantizers

We employ the linear quantizer for all weights and activations, except for the attention scores, which use a log2 quantizer to handle highly non-negative and uneven values [22, 49, 51]. For the the linear quantizer, given a full-precision input $\mathbf{x}$ and bit-width b, the quantized value $\mathbf{x}_q$ and the de-quantized value $\bar{\mathbf{x}}$ are computed as follows:

$$\mathbf{x}_q = \text{clip}\left(\left\lfloor \frac{\mathbf{x}}{\Delta} \right\rceil + z, 0, 2^b - 1\right), \bar{\mathbf{x}} = \Delta \cdot (\mathbf{x}_q - z), \quad (1)$$

where $\lfloor \cdot \rceil$ denotes rounding to the nearest integer, and clip limits the value to $[0, 2^b - 1]$. Here, Δ and z are the scale factor and zero-point, respectively. For the log2 quantizer:

$$\mathbf{x}_q = \text{clip}\left(\left\lfloor -\log_2 \frac{\mathbf{x}}{\Delta} \right\rceil, 0, 2^b - 1\right), \bar{\mathbf{x}} = \Delta \cdot 2^{-\mathbf{x}_q}. \quad (2)$$

3.1.2. Data Synthesis

DFQ methods parameterize synthetic images and optimize them toward real-like images with a pre-trained full-precision model F. Given a image $\tilde{\boldsymbol{I}}$ initialized from Gaussian noise, the one-hot loss [74] is introduced to learn label-related semantics:

$$\mathcal{L}^{\text{OH}}(\tilde{\boldsymbol{I}}) = CE(F(\tilde{\boldsymbol{I}}), c) \quad (3)$$

where $CE(\cdot, \cdot)$ represents the cross entropy, c is a random class label, and $F(\cdot)$ returns the predicted probability for image $\tilde{\boldsymbol{I}}$.

Moreover, total variance (TV) [76] loss is a smoothing regularization term to improve the image quality:

$$\mathcal{L}^{\text{TV}}(\tilde{\boldsymbol{I}}) = \iint |\nabla \tilde{\boldsymbol{I}}(\tau_1, \tau_2)| d\tau_1 d\tau_2. \quad (4)$$

where $\nabla \tilde{\boldsymbol{I}}(\tau_1, \tau_2)$ denotes the gradient at $\tilde{\boldsymbol{I}}$ at (τ_1, τ_2).

Classes \ Method	Real	PSAQ-ViT [47]	APA (Ours)
Class 1	0.68	0.44	0.64
Class 2	0.32	0.26	0.31
Class 3	0.41	0.31	0.36

Table 1. Average cosine similarity of three randomly selected classes. For real images, the similarity is measured within the class itself, while for PSAQ-ViT and APA, the similarity is measured between synthetic and real images of the same class. The results show that APA achieves higher similarity than PSE loss, indicating aligned semantics.

To perform DFQ for ViTs, PSAQ-ViT [47] proposes patch similarity entropy (PSE) loss. It first compute patch similarity $\mathbf{\Gamma}_l[i,j] = \frac{u_i \cdot u_j}{||u_i|| ||u_j||}$, where u_i, u_j are feature vectors of MHSA outputs in l-th block and $||\cdot||$ denotes the l_2 norm. Then, it estimates the density function $\hat{f}_l(x) = \frac{1}{Mh}\sum_{m=1}^{M} K\left(\frac{x-x_m}{h}\right)$, where $K(\cdot)$ is a normal kernel, h is the bandwidth, and x_m is the kernel center derived from $\mathbf{\Gamma}_l$. Finally, the PSE loss is defined as:

$$\mathcal{L}^{\text{PSE}}(\tilde{\boldsymbol{I}}) = \sum_{l=1}^{L} \int \hat{f}_l(x) \log\left[\hat{f}_l(x)\right] dx, \tag{5}$$

where L is the block number of the model.

3.2. Observations

Existing DFQ methods have made significant progress. However, through carefully analyzing the synthetic images, we reveal that these images suffer from issues of semantic distortion and semantic inadequacy, both of which hinder further advancement in the DFQ of ViTs.

The **semantic distortion** issue refers to the significant divergence between the semantics of synthetic images and real images. To demonstrate this, we visualize the features of synthetic and real images in Fig. 1a. Note that the penultimate feature typically is regarded to represent the semantics of the input [4, 58, 77, 78], It is clear that the features of PSAQ-ViT diverge significantly from real images, suggesting that these images fail to capture the true semantic distribution of real data. Tab. 1 quantitatively measures the semantics using the average cosine similarity (ranging from -1 to 1). We also report the intra-class similarity within real images as an approximate upper bound for comparison. This result further supports the observation of low similarity between synthetic images from PSAQ-ViT and real images. For instance, for Class 1, the intra-class similarity within real images is 0.68, while PSAQ-ViT achieves only 0.44.

The **semantic inadequacy** issue refers to the presence of dull regions in synthetic images, which contain redundant or non-semantic content [31, 35], hindering the model's learning process. As indicated in [13, 14], a diverse content and textures generally suggests rich information. As shown in Fig. 1b, many regions of synthetic images generated by PSAQ-ViT and PSAQ-ViT V2 exhibit a lack of diversity in content, with overly simplified textures. Specifically, the central region of PSAQ-ViT images only contains faint object structures, while PSAQ-ViT V2 images appear excessively smoothed and indistinct.

For high-bit quantization, where model capacity is largely retained [42], the performance degradation remains relatively minor even using semantic distorted and inadequate images [47, 62]. However, in low-bit quantization, where model capacity is severely damaged and informative images are essential for recovering performance [13, 83], fine-tuning on these poor-quality images leads to inferior generalization to real datasets, resulting in limited performance. For example, as shown in Tab. 2, the W4A4 ViT-B fine-tuned on real images yields 68.16%, whereas PSAQ-ViT only achieves 36.32%.

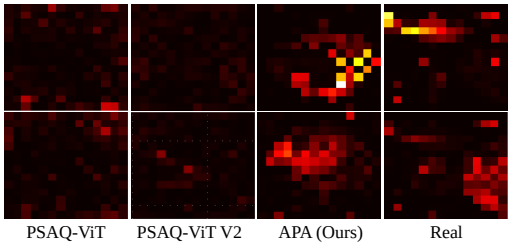


Figure 3. Comparison between attention maps.

3.3. Semantics Alignment and Reinforcement Data-Free Quantization

In the following, we introduce the proposed SARDFQ, whose framework is illustrated in Fig. 2.

3.3.1. Attention Priors Alignment

In ViTs, the self-attention mechanism encodes semantic correlations between image regions, where high-response areas in attention maps strongly correlate with semantic-discriminative content [10, 20, 29]. However, existing DFQ methods overlook this intrinsic property in the generation process. As a result, as shown in Fig. 3, synthetic images often exhibit disordered and unnatural attention patterns, with attention maps either overly diffuse or misaligned toward peripheral regions. This undermines their ability to preserve semantic-discriminative content, causing semantic distortion, as demonstrated in Fig. 1a and Tab. 1. In response, we propose Attention Priors Alignment (APA), which improves semantics alignment by optimizing synthetic images to follow randomly generated structure attention priors.

Specifically, given a synthetic image $\tilde{\boldsymbol{I}}$, we first obtain its attention maps in the h-th head of the l-th block, denoting as $\mathbf{A}_{l,h} \in \mathbb{R}^{N\times N}$, where N represents the total number of tokens. In DeiT, the attention of the classification token to-

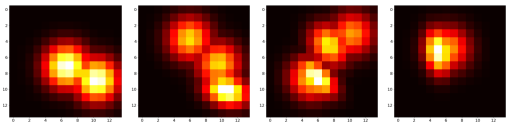


Figure 4. Examples of generated attention priors.

ward other tokens serves as the indicator for semantic versus non-semantic parts [5]. Thus, we extract $\mathbf{A}^c_{l,h} \in \mathbb{R}^{1\times(N-1)}$ from $\mathbf{A}_{l,h}$, representing the attention of the classification token to all tokens except itself. We then randomly generate attention priors $\tilde{\mathbf{A}}_{l,h}$, whose generation is detailed in the next part, and align $\mathbf{A}^c_{l,h}$ with $\tilde{\mathbf{A}}_{l,h}$ by:

$$\mathcal{L}_{l,h}(\tilde{\boldsymbol{I}}) = \text{MSE}(\mathbf{A}^c_{l,h} - \tilde{\mathbf{A}}_{l,h}), \tag{6}$$

where MSE represents the mean squared error. For Swin models that do not use a classification token, we substitute $\mathbf{A}^c_{l,h}$ in Eq. 6 with the average attention map of all tokens [10]. As noted in [17], ViTs initially focus on all regions to capture low-level information in shallow blocks and gradually shift their focus toward semantic regions in deeper blocks to extract high-level semantic information. Leveraging this property, we selectively apply $\mathcal{L}^{\text{APA}}_{l,h}$ to deeper blocks, progressively aligning attention towards semantically relevant areas. The total APA loss is computed as a depth-weighted sum of the individual Eq. 6 across these deeper blocks:

$$\mathcal{L}^{\text{APA}}(\tilde{\boldsymbol{I}}) = \sum_{l=S}^{L}\sum_{h=1}^{H}\frac{l}{L}\mathcal{L}_{l,h}(\tilde{\boldsymbol{I}}), \tag{7}$$

where S is a pre-given hyper-parameter denoting the start of deep blocks and are experimentally set $S = \frac{L}{2}$.

Attention Priors Generation. To generate attention priors $\tilde{\mathbf{A}}_{l,h}$, Gaussian Mixture Models (GMMs) are employed as it is the most commonly used distribution with high flexibility. ViTs utilize different attention heads to capture diverse patterns and learn varied, informative representations [8]. Thus, we use distinct GMMs for each head. Note that the goal here is not to replicate real attention maps precisely, but to generate simulated structure attention priors to guide the learning of synthetic images.

In particular, we first initialize an all zero matrix $\tilde{\mathbf{P}} \in \mathbb{R}^{H\times W}$, where $H = W = \sqrt{N-1}$ for DeiT, $H = W = \sqrt{N}$ for Swin. For example, for DeiT-S, the $H = W = \sqrt{196} = 14$. Then, we generate k two-dimensional Gaussian distributions, where k is randomly sampled from $1 \sim K_{APA}$ and K_{APA} is set to 5 in all experiments. Each Gaussian has its mean and covariance[4]. Consequently, the matrix element at the i-th row and j-th column, $\tilde{\mathbf{P}}[i,j]$, is determined by:

$$\tilde{\mathbf{P}}[i,j] = \max_{m=1,\cdots,k} \mathbf{G}^m[i,j]. \tag{8}$$

[4]The pseudo code is detailed in appendix.

Then, $\tilde{\mathbf{P}}$ is normalized by:

$$\tilde{\mathbf{P}}_n = \frac{\tilde{\mathbf{P}}}{\sum\tilde{\mathbf{P}}}\cdot(1-x). \tag{9}$$

Here, for DeiT that incorporates the classification token, x is randomly sampled from a uniform distribution $U(0,1)$, representing the proportion of the attention score that the classification token allocates to itself. For Swin, which does not use a classification token, x is set to 0. Finally, $\tilde{\mathbf{P}}_n$ is flatten to match the dimensionality:

$$\tilde{\mathbf{A}}_{l,h} = \text{flatten}(\tilde{\mathbf{P}}_n). \tag{10}$$

Fig. 4 displays examples of generated attention priors.

Discussion. APA prevents attention disorder, ensuring that synthetic images exhibit more coherent and natural attention patterns, as demonstrated in Fig. 3. As a result, APA selectively enhances the responses in certain regions, effectively emphasizing semantic-discriminative regions within synthetic images and thus prompting the discriminative features. Although simple, APA enables synthetic images to align better with the real semantics, as validated by both visual and quantitative evaluations. As shown in Fig. 1a, compared to PSAQ-ViT, features obtained after applying the APA loss are more closely aligned with real images, indicating better semantic alignment. The quantitative results in Tab. 1 further support that APA achieves superior semantic alignment. For example, in Class 1, the intra-class similarity for PSAQ-ViT is 0.44, whereas APA achieves a higher similarity of 0.64.

3.3.2. Multi-Semantic Reinforcement

Current DFQ methods for ViTs [47, 48, 62] optimize images through global optimization, treating the entire image as a single semantic unit. However, this is affected by low-rank structural regularity [32], where adjacent pixels exhibit strong similarity, leading to dull regions with redundant or non-semantic content for model learning [35]. This issue is further exacerbated by the tokenization mechanism in ViTs, as processing images in fixed-size patches make dull regions increase at the patch level [19, 69]. Consequently, as shown in Fig. 1b, synthetic images generated by existing methods exhibit large dull regions, resulting in semantic inadequacy [35]. In response, we propose Multi-Semantic Reinforcement (MSR), which applies localized patch optimization to enhance semantic richness by learning local patches with distinct semantics.

Specifically, for a synthetic image $\tilde{\boldsymbol{I}}$, instead of feeding only the entire image, we also feed its patches and optimize them individually. Initially, we select m non-overlapping patches, where m is chosen randomly from the set $\{1, 2, \ldots, K_{MSR}\}$, with K_{MSR} set to 4 in all experiments. These m patches are then cropped and resized to

match the model's input dimensions:

$$\{\tilde{\boldsymbol{I}}^i_{MSR}\}_{i=1,\cdots,m} = \text{resize}\big(\text{crop}_m(\tilde{\boldsymbol{I}})\big), \tag{11}$$

where $\text{crop}_m(\cdot)$ crops m non-overlapping patches from input, and $\text{resize}(\cdot)$ is the resize function. Each patch, denoted as $\tilde{\boldsymbol{I}}^i_{MSR}$, is treated as a new image with an assigned semantic target c^i. Note that the gradient is backpropagated only to update the corresponding patch in the original image, leaving the rest of the image unaffected.

Softlabel Learning. The one-hot loss (Eq. 3) only learns the semantics of the target class, making it unsuitable for $\tilde{\boldsymbol{I}}$ under MSR, as its patches $\{\tilde{\boldsymbol{I}}^i_{MSR}\}_{i=1,\cdots,m}$ contain distinct semantics. In response, we propose Softlabel Learning (SL), which applies multiple semantic targets to accommodate the learning of images augmented by MSR. Specifically, we first sample $Z \in \mathbb{R}^C \sim U(0,1)$, then modify its values by:

$$\begin{cases} Z[c^i] \sim U(\epsilon_1, \epsilon_2), & \text{for } \tilde{\boldsymbol{I}}^i_{MSR}, \\ Z[c^1, \ldots, c^m] \sim U(\epsilon_1, \epsilon_2), & \text{for } \tilde{\boldsymbol{I}}, \end{cases}$$

where $U(\epsilon_1, \epsilon_2)$ denotes the uniform distribution over the interval $[\epsilon_1, \epsilon_2]$, m is the number of patches determined in MSR, and ϵ_1 and ϵ_2 control the softness, both empirically set consistently to 5 and 10 in all experiments. The soft target is defined as $T_s = \text{softmax}(Z)$, and SL loss is:

$$\mathcal{L}^{\text{SL}}(\tilde{\boldsymbol{I}}/\tilde{\boldsymbol{I}}^i_{MSR}) = SCE\big(F(\tilde{\boldsymbol{I}}/\tilde{\boldsymbol{I}}^i_{MSR}), T_s\big), \tag{12}$$

where $SCE(\cdot,\cdot)$ is soft cross entropy and $F(\cdot)$ returns the predicted probability for its input. SL facilitates smooth learning between across semantic targets, ensuring that MSR-enhanced images receive consistent supervision rather than conflicting supervision.

Discusion. By leveraging localized patch optimization, MSR ensures that each patch contributes unique semantics, forcing synthetic images to capture diverse features rather than being dominated by large homogeneous dull regions. As demonstrated by the richer and more diverse content and textures in Fig. 1b, MSR effectively reduces dull regions and enhances semantic richness in synthetic images. Moreover, MSR transforms synthetic images into composites of multiple semantic objects rather than a single unit, thereby providing more distinct semantic samples, *i.e.*, $\{\tilde{\boldsymbol{I}}^i_{MSR}\}_{i=1,\cdots,m}$, for model training. Unlike traditional cropping used in data augmentation, which aims to improve classification robustness by training with cropped patches labeled with the original class, MSR aims for semantic richness within synthetic images, ultimately enabling accurate data-free quantization.

3.4. Overall Pipeline

The overall pipeline consists of two stages: data synthesis and quantized network learning. The first stage uses the proposed SARDFQ to produce synthetic images. The second stage fine-tunes the quantized model using the generated synthetic images.

3.4.1. Data Synthesis

In the data synthesis stage, we combine the proposed APA loss of Eq. 7, SL loss of Eq. 12, and TV loss of Eq. 4 to formulate the objective function as follows:

$$\mathcal{L}_G(\tilde{\boldsymbol{I}}) = \alpha_1 \mathcal{L}^{\text{APA}}(\tilde{\boldsymbol{I}}) + \mathcal{L}^{\text{SL}}(\tilde{\boldsymbol{I}}) + 0.05\mathcal{L}^{\text{TV}}(\tilde{\boldsymbol{I}}). \tag{13}$$

where α_1 is hyperparameters and is determined by grid-search. Note that the weight of TV loss is fixed to 0.05, following [47], to avoid a cumbersome hyperparameter search.

3.4.2. Quantized Network Learning

Recently DFQ methods have introduced the PTQ methods in learning quantized models due to their advantages of speed, memory efficiency, and performance [33, 62]. Thus, following the success of [42, 70], we fine-tune the quantized network block-wisely. Specifically, denote $\mathbf{X}_l$ as the outputs of the l-th block of the full-precision model, and $\bar{\mathbf{X}}_l$ represent outputs of the quantized counterpart. The reconstruction loss is defined as:

$$\mathcal{L}_l = \|\mathbf{X}_l - \bar{\mathbf{X}}_l\|_2. \tag{14}$$

Here, $\mathcal{L}_l$ is only backward to update weights within l-th block. Note that for a fair comparison, all compared methods adopt the same quantized network learning stage.

4. Experiment

4.1. Implementation Details

Models and Tasks. We evaluate the performance of SARDFQ by test quantized ViT-S/B [20], DeiT-T/S/B [67], and Swin-S/B [54] on the classification task using ImageNet [63]. The pre-trained models are downloaded from the timm library. In appendix, we further provide results on detection and segmentation tasks.

Comparison methods. We compare our SARDFQ against Gaussian noise, real images, and previous methods including SMI [31] and PSAQ-ViT [47] and its subsequent version, PSAQ-ViT V2 [48]. For a fair comparison, we generate synthetic images using their methods and apply our quantized network learning strategy. We use the official code for SMI and PSAQ-ViT to reproduce their images, while PSAQ-ViT V2 is re-implemented by us, as no official code is available.[5]

Experimental settings. All experiments were conducted using the PyTorch framework [60] on a single NVIDIA 3090 GPU. In the data synthesis stage, synthetic

[5] In appendix, we present practical efficiency results and more performance comparisons with more methods such as CLAMP-ViT [62] on W8/A8 and W4/A8 settings.

Model	W/A	Real	Gaussian noise	PSAQ-ViT [47]	PSAQ-ViT V2 [48]	SMI [31]	SARDFQ (Ours)
ViT-S (81.39)	4/4	66.57	6.02	47.24	41.53	$24.33_{29.41}$	**50.32**
	5/5	76.69	36.77	71.59	68.41	$61.33_{65.19}$	**74.31**
	6/6	79.46	61.20	77.20	74.76	$72.95_{72.46}$	**78.40**
ViT-B (84.54)	4/4	68.16	0.15	36.32	26.32	$35.27_{19.67}$	**51.84**
	5/5	79.21	4.16	68.48	67.95	$67.53_{57.13}$	**70.70**
	6/6	81.89	55.18	76.65	71.87	$76.33_{69.82}$	**79.16**
DeiT-T (72.21)	4/4	56.60	17.43	47.75	30.20	$30.14_{13.18}$	**52.06**
	5/5	67.09	43.49	64.10	55.16	$56.44_{39.35}$	**66.41**
	6/6	69.81	56.23	68.37	62.77	$64.03_{44.39}$	**69.73**
DeiT-S (79.85)	4/4	68.46	20.89	58.28	45.53	$42.77_{11.71}$	**62.29**
	5/5	75.06	41.06	71.90	63.14	$62.88_{29.13}$	**74.06**
	6/6	77.87	65.63	75.85	68.85	$71.65_{37.69}$	**77.31**
DeiT-B (81.85)	4/4	77.07	47.20	71.75	66.43	$65.33_{59.04}$	**72.17**
	5/5	79.86	65.46	78.45	76.77	$76.74_{75.33}$	**78.72**
	6/6	80.90	62.79	80.00	79.22	$78.81_{77.66}$	**80.15**
Swin-S (83.20)	4/4	78.12	31.92	73.19	65.55	65.85	**74.74**
	5/5	80.51	52.10	78.15	74.37	75.41	**79.56**
	6/6	80.60	65.66	79.74	78.50	78.25	**80.56**
Swin-B (85.27)	4/4	78.80	30.14	71.84	67.42	65.23	**76.42**
	5/5	82.51	35.28	78.50	77.20	75.25	**80.82**
	6/6	82.64	67.37	82.00	81.41	80.30	**83.03**

Table 2. Quantization results on ImageNet dataset, with top-1 accuracy (%) reported. The performance of the full-precision model is listed below the model name. "W/A" denotes the bit-width of weights/activations. "Real" refers to using real images. For SMI [31], we provide the performance of using dense (normal-sized numbers) and sparse (smaller-sized numbers) synthetic images, respectively. Note that for Swin models, we do not provide the results for sparse synthetic images as the sparse generation method of SMI is infeasible.

images were initialized with standard Gaussian noise, generating 32 images in total. The Adam optimizer [38] with $\beta_1 = 0.5$, $\beta_2 = 0.9$ was used, with learning rates of 0.25 for Swin and 0.2 for others, and a total of 1,000 iterations. For all models, K_{APA}, K_{MSR}, ϵ_1, and ϵ_2 were set to 5, 4, 5, and 10, respectively, based on a search with W4A4 DeiT-S. The value of α_1 was determined by grid search for each model: 1e5 for DeiT-T/S, 1e4 for DeiT-B, 100 for ViT-B and Swin-B, 10 for Swin-S, and 1 for ViT-S. Although further hyperparameter search may improve performance, the current settings already yield superior results. In the quantized network learning stage, following [84], the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ was used, with weight decay set to 0 and an initial learning rate of 4e-5, adjusted via cosine decay for 100 iterations. A channel-wise quantizer was used for weights, and a layer-wise quantizer for activations, with all matrix multiplications in ViTs quantized [47–49].

4.2. Quantization Results

The quantization results are presented in Tab. 2. Our SARDFQ demonstrates consistent improvements across various quantization bit-width configurations, particularly with low bit-width settings. Specifically, for ViT-S, SARDFQ improves the performance by 3.08% in the W4/A4 setting, 2.72% in the W5/A5 setting, and 1.20% in the W6/A6 setting. For ViT-B, SARDFQ achieves performance gains of 15.52% in the W4/A4 setting, 2.22% in the W5/A5 setting, and 2.51% in the W6/A6 setting. Results on DeiT also demonstrate the effectiveness of SARDFQ. For example, on DeiT-T, SARDFQ shows a marked improvement by increasing top-1 accuracy by 4.31% in the W4/A4 setting, 2.31% in the W5/A5 setting, and 1.36% in the W6/A6 setting. For DeiT-S, SARDFQ enhances top-1 accuracy by 4.01% in the W4/A4 setting, 2.16% in the W5/A5 setting, and 1.46% in the W6/A6 setting. The quantization results of Swin-S/B also affirm the superiority of our SARDFQ in enhancing model accuracy under different quantization configurations. In particular, for Swin-S, the proposed SARDFQ increases the accuracy by 1.55% for the W4/A4 setting, 1.41% for the W5/A5 setting, and 0.82% for the W6/A6 setting, respectively. When it comes to Swin-B, the proposed SARDFQ increases the accuracy by 4.58% for the W4/A4 setting, 2.32% for the W5/A5 setting, and 1.03%

for the W6/A6 setting, respectively.

4.3. Ablation Study

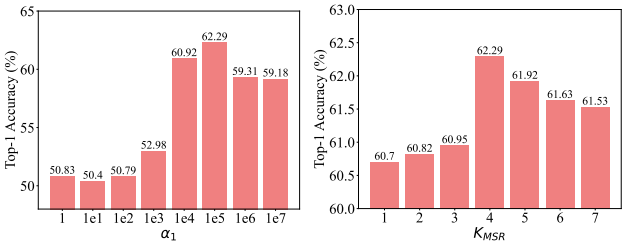
All ablation studies are conducted on the W4A4 DeiT-S.

APA	MSR	SL	Acc. (%)
Baseline			51.73
✓			60.26
	✓		50.75
		✓	52.02
✓	✓		61.58
✓		✓	60.51
	✓	✓	56.08
✓	✓	✓	**62.29**

Table 3. Influence of the proposed APA, MSR, and SL on accuracy. The baseline adopts the one-hot loss.

Analysis of APA, MSR, and SL. We analyze the effectiveness of the proposed APA (Sec.,3.3.1), MSR (Sec.,3.3.2), and SL (Eq.,12) in Tab.,3. Adding APA and SL individually to the baseline increases accuracy. Notably, APA boosts performance from 51.73% to 60.26%, confirming its effectiveness in aligning semantics (Sec.,3.3.1). Applying MSR alone slightly decreases accuracy from 51.73% to 50.75%, indicating that one-hot loss is unsuitable for MSR-augmented synthetic images. However, when both MSR and SL are applied, accuracy rises to 56.08%, suggesting SL is more compatible with MSR than one-hot loss. Combining APA with either MSR or SL further improves performance. For example, APA and SL together yield an accuracy of 60.51%, and when all three strategies are used, the best performance of 62.29% is achieved.

Analysis of priors distribution. Tab. 4a showcases the results of using other distributions to formulate the attention priors. The unevenly distributed GMM and Laplace present comparable performance of 62.29% and 62.16%, respectively. Moreover, GMM provides a similar performance to the real's, indicating it performs well in imitating the patterns of real images.



(a) (b)

Figure 5. Effect of varying (a) α_1 and (b) K_{MSR}.

Priors Distribution	Top-1
GMM	62.29
Laplace	62.16
Real	**63.19**

(a)

K_{APA}	Top-1
1	61.13
3	61.52
5	**62.29**
7	61.53
9	61.05

(b)

S	Top-1
0	61.96
$L/2$	**62.29**

(c)

w. $\frac{l}{L}$	Top-1
✓	**62.29**
×	61.32

(d)

Table 4. Effect of varying (a) priors types; (b) K_{APA}; (c) S; (d) $\frac{l}{L}$.

Analysis of α_1, K_{APA}, and K_{MSR}. The α_1 from Eq. 13 balance the importance of the proposed APA loss during the update of the synthetic images. Fig. 5a demonstrates that the optimal performance is achieved when $\alpha_1 = 1e5$. Incrementally increasing α_1 improves performance up to 62.29% at $\alpha_1 = 1e5$. However, further increases in α_1 subsequently degrade performance. The K_{APA} in APA is the upper limit on the number of the Gaussian distributions used for priors generation. Tab. 4b displays the ablation study for different values of K_{APA}. The best accuracy is achieved when $K_{APA} = 5$. The K_{MSR} in MSR is the upper limit on the number of patches. Fig. 5b demonstrates that when $K_{MSR} = 4$, the optimal performance is achieved. Using K_{MSR} larger than 4 will hurt the accuracy. We consider this due to limited patch resolution if using a too large K_{MSR}.

Analysis of APA loss. Here, we conduct the ablation study by considering the S and scale $\frac{l}{L}$ in Eq. 7. Tab. 4c presents the effect of varying S in Eq. 7. If applying APA loss to all blocks ($S = 0$), the top-1 accuracy decreases to 61.96%. From Tab. 4d, it can be seen that absorbing the scale $\frac{l}{L}$ in Eq. 7 presents 0.97% performance gains.

5. Limitations

We further discuss some limitations of the proposed SARDFQ, which will guide future research directions. First, although SARDFQ shows substantial performance improvement, a performance gap between SARDFQ and real data remains challenging, highlighting the need for a stronger semantics alignment and reinforcement method. Second, SARDFQ currently lacks a theoretical foundation. Future work could establish a theoretical framework for SARDFQ, particularly in understanding how APA and MSR influence synthetic images in a formalized manner.

6. Conclusion

In this paper, we investigate the DFQ method for ViTs. We first identify that synthetic images generated by existing methods suffer from semantic distortion and inadequacy issues, and propose SARDFQ to address these issues. To mitigate semantic distortion, SARDFQ introduces APA, which guides synthetic images to align with randomly generated structural attention patterns. To tackle semantic inadequacy, SARDFQ incorporates MSR. MSR optimizes different regions of synthetic images with unique semantics, thereby enhancing overall semantic richness. Moreover, SARDFQ employs SL, which adopts multiple semantic targets to ensure seamless learning of images augmented by MSR. Extensive experiments on various ViT models and tasks validate the effectiveness of SARDFQ.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *IEEE/CVF international conference on computer vision (ICCV)*, pages 6836–6846, 2021. 2

[2] 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. *arXiv preprint arXiv:2310.03661*, 2023. 3

[3] Ron Banner, Yury Nahshan, Daniel Soudry, et al. Post training 4-bit quantization of convolutional networks for rapid-deployment. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 7950–7958, 2019. 3

[4] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(8):1798–1828, 2013. 4

[5] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 5

[6] Yaohui Cai, Zhewei Yao, Zhen Dong, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Zeroq: A novel zero shot quantization framework. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13169–13178, 2020. 2, 3

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, pages 213–229. Springer, 2020. 2

[8] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 397–406, 2021. 5

[9] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12299–12310, 2021. 2

[10] Mengzhao Chen, Mingbao Lin, Ke Li, Yunhang Shen, Yongjian Wu, Fei Chao, and Rongrong Ji. Cf-vit: A general coarse-to-fine method for vision transformer. In *AAAI Conference on Artificial Intelligence*, pages 7042–7052, 2023. 4, 5

[11] Mengzhao Chen, Wenqi Shao, Peng Xu, Mingbao Lin, Kaipeng Zhang, Fei Chao, Rongrong Ji, Yu Qiao, and Ping Luo. Diffrate: Differentiable compression rate for efficient vision transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17164–17174, 2023. 2

[12] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8126–8135, 2021. 2

[13] Xinrui Chen, Yizhi Wang, Renao Yan, Yiqing Liu, Tian Guan, and Yonghong He. Texq: Zero-shot network quantization with texture feature distribution calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3, 4

[14] Kanghyun Choi, Deokki Hong, Noseong Park, Youngsok Kim, and Jinho Lee. Qimera: Data-free quantization with synthetic boundary supporting samples. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 14835–14847, 2021. 2, 4

[15] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8311–8321, 2022. 3

[16] Kanghyun Choi, Hye Yoon Lee, Dain Kwon, SunJong Park, Kyuyeun Kim, Noseong Park, and Jinho Lee. Mimiq: Low-bit data-free quantization of vision transformers. *arXiv preprint arXiv:2407.20021*, 2024. 2, 3

[17] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. In *International Conference on Learning Representations (ICLR)*, 2020. 5

[18] Yifu Ding, Haotong Qin, Qinghua Yan, Zhenhua Chai, Junjie Liu, Xiaolin Wei, and Xianglong Liu. Towards accurate post-training quantization for vision transformer. In *30th ACM International Conference on Multimedia (ACMMM)*, pages 5380–5388, 2022. 3

[19] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International conference on machine learning (ICML)*, pages 2793–2803. PMLR, 2021. 5

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 2, 4, 6

[21] Steven K. Esser, Jeffrey L. McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S. Modha.

Learned step size quantization. In *International Conference on Learning Representations (ICLR)*, 2020. 3

[22] Natalia Frumkin, Dibakar Gope, and Diana Marculescu. Jumping through local minima: Quantization in the loss landscape of vision transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16978–16988, 2023. 3

[23] Ruihao Gong, Xianglong Liu, Shenghu Jiang, Tianxiang Li, Peng Hu, Jiazhen Lin, Fengwei Yu, and Junjie Yan. Differentiable soft quantization: Bridging full-precision and low-bit neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4852–4861, 2019. 3

[24] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2672–2680, 2014. 3

[25] 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 *The Eleventh International Conference on Learning Representations (ICLR)*, 2022. 2, 3

[26] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[27] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(1):87–110, 2022. 2

[28] Zhiwei Hao, Jianyuan Guo, Ding Jia, Kai Han, Yehui Tang, Chao Zhang, Han Hu, and Yunhe Wang. Learning efficient vision transformers via fine-grained manifold distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[29] Joakim Bruslund Haurum, Sergio Escalera, Graham W Taylor, and Thomas B Moeslund. Which tokens to use? investigating token reduction in vision transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 773–783, 2023. 4

[30] Zejiang Hou and Sun-Yuan Kung. Multi-dimensional vision transformer compression via dependency guided gaussian process search. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3669–3678, 2022. 2

[31] 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 *International Conference on Machine Learning (ICML)*, 2024. 2, 3, 4, 6, 7

[32] Jinggang Huang and David Mumford. Statistics of natural images and models. In *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 541–547. IEEE, 1999. 5

[33] Yongkweon Jeon, Chungman Lee, and Ho-young Kim. Genie: show me the data for quantization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12064–12073, 2023. 3, 6

[34] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*, 2021. 2

[35] Jang-Hyun Kim, Jinuk Kim, Seong Joon Oh, Sangdoo Yun, Hwanjun Song, Joonhyun Jeong, Jung-Woo Ha, and Hyun Oh Song. Dataset condensation via efficient synthetic-data parameterization. In *International Conference on Machine Learning (ICML)*, pages 11102–11118. PMLR, 2022. 4, 5

[36] Minjun Kim, Jaehyeon Choi, Jongkeun Lee, Wonjin Cho, and U Kang. Zero-shot quantization: A comprehensive survey. In *Thirty-four International Joint Conference on Artificial Intelligence, IJCAI*, 2025. 3

[37] Minjun Kim, Jongjin Kim, and U Kang. Synq: Accurate zero-shot quantization by synthesis-aware fine-tuning. In *International Conference on Learning Representations (ICLR)*, 2025. 3

[38] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2014. 7

[39] Raghuraman Krishnamoorthi. Quantizing deep convolutional networks for efficient inference: A whitepaper. *arXiv preprint arXiv:1806.08342*, 2018. 2

[40] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24417–24426, 2023. 3

[41] Yuhang Li, Xin Dong, and Wei Wang. Additive powers-of-two quantization: An efficient non-uniform discretization for neural networks. In *International Conference on Learning Representations (ICLR)*, 2020. 3

[42] 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 *International Conference on Learning Representations (ICLR)*, 2021. 3, 4, 6

[43] Yanjing Li, Sheng Xu, Baochang Zhang, Xianbin Cao, Peng Gao, and Guodong Guo. Q-vit: Accurate and fully quantized low-bit vision transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 34451–34463, 2022. 3

[44] Yuhang Li, Youngeun Kim, Donghyun Lee, and Priyadarshini Panda. Stableq: Enhancing data-scarce quantization with text-to-image data. *arXiv preprint arXiv:2312.05272*, 2023. 3

[45] Yuhang Li, Youngeun Kim, Donghyun Lee, Souvik Kundu, and Priyadarshini Panda. Genq: Quantization in low data regimes with generative synthetic data. In *European Conference on Computer Vision (ECCV)*, pages 216–235. Springer, 2024. 3

[46] Zhikai Li and Qingyi Gu. I-vit: Integer-only quantization for efficient vision transformer inference. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17065–17075, 2023. 2, 3

[47] Zhikai Li, Liping Ma, Mengjuan Chen, Junrui Xiao, and Qingyi Gu. Patch similarity aware data-free quantization for vision transformers. In *European Conference on Computer Vision (ECCV)*, pages 154–170. Springer, 2022. 1, 2, 3, 4, 5, 6, 7

[48] Zhikai Li, Mengjuan Chen, Junrui Xiao, and Qingyi Gu. Psaq-vit v2: Toward accurate and general data-free quantization for vision transformers. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2023. 2, 3, 5, 6, 7

[49] Zhikai Li, Junrui Xiao, Lianwei Yang, and Qingyi Gu. Repq-vit: Scale reparameterization for post-training quantization of vision transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17227–17236, 2023. 3, 7

[50] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *IEEE/CVF international conference on computer vision (ICCV)*, pages 1833–1844, 2021. 2

[51] Yang Lin, Tianyu Zhang, Peiqin Sun, Zheng Li, and Shuchang Zhou. Fq-vit: Post-training quantization for fully quantized vision transformer. In *Thirty-First International Joint Conference on Artificial Intelligence, (IJCAI)*, pages 1173–1179, 2022. 3

[52] Shih-Yang Liu, Zechun Liu, and Kwang-Ting Cheng. Oscillation-free quantization for low-bit vision transformers. In *International Conference on Machine Learning (ICML)*, pages 21813–21824, 2023. 3

[53] Yijiang Liu, Huanrui Yang, Zhen Dong, Kurt Keutzer, Li Du, and Shanghang Zhang. Noisyquant: Noisy bias-enhanced post-training activation quantization for vision transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20321–20330, 2023. 3

[54] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF international conference on computer vision (ICCV)*, pages 10012–10022, 2021. 2, 6

[55] Zhenhua Liu, Yunhe Wang, Kai Han, Wei Zhang, Siwei Ma, and Wen Gao. Post-training quantization for vision transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 28092–28103, 2021. 2, 3

[56] Yiwei Ma, Jiayi Ji, Xiaoshuai Sun, Yiyi Zhou, and Rongrong Ji. Towards local visual modeling for image captioning. *Pattern Recognition*, 138:109420, 2023. 2

[57] Sachin Mehta and Mohammad Rastegari. Mobilevit: Lightweight, general-purpose, and mobile-friendly vision transformer. In *International Conference on Learning Representations (ICLR)*, 2022. 2

[58] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 23296–23308, 2021. 4

[59] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3163–3172, 2021. 2

[60] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8026–8037, 2019. 6

[61] Biao Qian, Yang Wang, Richang Hong, and Meng Wang. Adaptive data-free quantization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7960–7968, 2023. 3

[62] Akshat Ramachandran, Souvik Kundu, and Tushar Krishna. Clamp-vit: contrastive data-free learning for adaptive post-training quantization of vits. In *European Conference on Computer Vision (ECCV)*. Springer, 2024. 2, 3, 4, 5, 6

[63] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115:211–252, 2015. 6

[64] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. Transformers in medical imaging: A survey. *Medical Image Analysis*, page 102802, 2023. 2

[65] Huixin Sun, Runqi Wang, Yanjing Li, Xianbin Cao, Xiaolong Jiang, Yao Hu, and Baochang Zhang. P4q: Learning to prompt for quantization in visual-language models. *arXiv preprint arXiv:2409.17634*, 2024. 3

[66] Yehui Tang, Kai Han, Yunhe Wang, Chang Xu, Jianyuan Guo, Chao Xu, and Dacheng Tao. Patch slimming for efficient vision transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12165–12174, 2022. 2

[67] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*, pages 10347–10357. PMLR, 2021. 2, 6

[68] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research (JMLR)*, 9:2579–2605, 2008. 1

[69] Peihao Wang, Wenqing Zheng, Tianlong Chen, and Zhangyang Wang. Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. *arXiv preprint arXiv:2203.05962*, 2022. 5

[70] Xiuying Wei, Ruihao Gong, Yuhang Li, Xianglong Liu, and Fengwei Yu. Qdrop: Randomly dropping quantization for extremely low-bit post-training quantization. In *International Conference on Learning Representations (ICLR)*, 2022. 6

[71] Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and improving relative position encoding for vision transformer. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10033–10041, 2021. 2

[72] Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Tinyvit: Fast pretraining distillation for small vision transformers. In *European Conference on Computer Vision (ECCV)*, pages 68–85, 2022. 2

[73] Zhiqiang Shen Xijie Huang and Kwang-Ting Cheng. Variation-aware vision transformer quantization. *arXiv preprint arXiv:2307.00331*, 2023. 3

[74] Shoukai Xu, Haokun Li, Bohan Zhuang, Jing Liu, Jiezhang Cao, Chuangrun Liang, and Mingkui Tan. Generative low-bitwidth data free quantization. In *European Conference on Computer Vision (ECCV)*, pages 1–17, 2020. 3

[75] Shoukai Xu, Haokun Li, Bohan Zhuang, Jing Liu, Jiezhang Cao, Chuangrun Liang, and Mingkui Tan. Generative low-bitwidth data free quantization. In *European Conference on Computer Vision (ECCV)*, pages 1–17. Springer, 2020. 2

[76] 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 transffer via deepinversion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8715–8724, 2020. 2, 3

[77] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015. 4

[78] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 818–833, 2014. 4

[79] Jinnian Zhang, Houwen Peng, Kan Wu, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Minivit: Compressing vision transformers with weight multiplexing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12145–12154, 2022. 2

[80] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15658–15667, 2021. 2, 3

[81] Dehua Zheng, Wenhui Dong, Hailin Hu, Xinghao Chen, and Yunhe Wang. Less is more: Focus attention for efficient detr. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6674–6683, 2023. 2

[82] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 6881–6890, 2021. 2

[83] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12339–12348, 2022. 2, 4

[84] Yunshan Zhong, Jiawei Hu, Mingbao Lin, Mengzhao Chen, and Rongrong Ji. I&s-vit: An inclusive & stable method for pushing the limit of post-training vits quantization. *arXiv preprint arXiv:2311.10126*, 2023. 7

[85] Yunshan Zhong, Jiawei Hu, You Huang, Yuxin Zhang, and Rongrong Ji. Erq: Error reduction for post-training quantization of vision transformers. In *International Conference on Machine Learning (ICML)*, 2024. 3