MimiQ: Low-Bit Data-Free Quantization of Vision Transformers with Encouraging Inter-Head Attention Similarity

Kanghyun Choi[1], **Hyeyoon Lee**[1], **Dain Kwon**[1], **SunJong Park**[1], **Kyuyeun Kim**[2], **Noseong Park**[3], **Jonghyun Choi**[1], **Jinho Lee**[1]*

[1]Seoul National University
[2]Google
[3]KAIST

{kanghyun.choi, hylee817, dain.kwon, ryan0507, jonghyunchoi, leejinho}@snu.ac.kr,
kyuyeunk@google.com, noseong@kaist.ac.kr

Abstract

Data-free quantization (DFQ) is a technique that creates a lightweight network from its full-precision counterpart without the original training data, often through a synthetic dataset. Although several DFQ methods have been proposed for vision transformer (ViT) architectures, they fail to achieve efficacy in low-bit settings. Examining the existing methods, we observe that their synthetic data produce misaligned attention maps, while those of the real samples are highly aligned. From this observation, we find that aligning attention maps of synthetic data helps improve the overall performance of quantized ViTs. Motivated by this finding, we devise MimiQ, a novel DFQ method designed for ViTs that enhances inter-head attention similarity. First, we generate synthetic data by aligning head-wise attention outputs from each spatial query patch. Then, we align the attention maps of the quantized network to those of the full-precision teacher by applying head-wise structural attention distillation. The experimental results show that the proposed method significantly outperforms baselines, setting a new state-of-the-art for ViT-DFQ.

Code — https://github.com/iamkanghyunchoi/mimiq
Extended version — https://arxiv.org/abs/2407.20021

Introduction

Over the past few years, Vision Transformers (ViT) (Dosovitskiy et al. 2021) have gained increasing interest due to their remarkable performance on many computer vision tasks. However, ViT has high computational costs compared to conventional CNNs, making it challenging to adopt in many resource-constrained devices. Thus, various works focus on reducing the costs of ViT architectures (Li et al. 2022a; Liu et al. 2021c; Kong et al. 2022; Yu and Wu 2023). One popular approach is network quantization (Nagel et al. 2021; Gholami et al. 2021), which converts floating-point parameters and features to low-bit integers. However, naively converting the parameters to lower-bit induces a large accuracy drop, which is why quantization usually requires additional calibration (Sung, Shin, and Hwang 2015; Liu et al. 2021c; Lin et al. 2022) or fine-tuning (Courbariaux, Bengio, and David 2015; Hubara et al. 2017) using the original training dataset. Unfortunately, in real-life cases, the original training dataset is not always available due to privacy concerns, security issues, or copyright protections (Liu et al. 2021a; Hathaliya and Tanwar 2020).

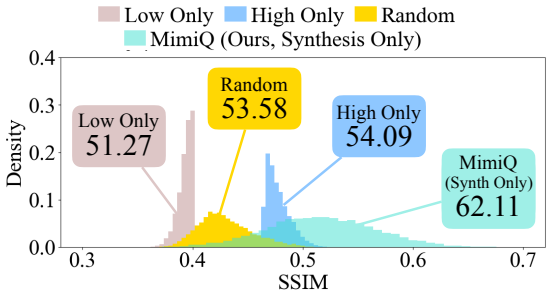


Figure 1: Attention similarity histograms of Base DF Synthesis (high and low similarity, random sampled) and MimiQ. Colored boxes denote ImageNet accuracy of the corresponding dataset. This motivational study shows the attention similarity is related to the DFQ accuracy of ViTs.

Data-free quantization (DFQ) (Nagel et al. 2019) addresses such dataset inaccessibility by quantizing the network without using the original training data. To replace the original dataset, recent methods generate synthetic data from the pretrained networks and use it for calibration. These approaches directly optimize synthetic samples with gradient descent (Zhong et al. 2022; Li et al. 2022b, 2023a) or train an auxiliary data generator (Xu et al. 2020; Choi et al. 2021).

Unfortunately, existing DFQ methods suffer from destructive accuracy drops on low-bit ViTs (refer to Tab. 2). This is because CNN-based DFQs rely on batch normalization (BN) statistics to create synthetic samples resembling the original training data, making them unsuitable for ViTs that do not contain the BN layer. Recent DFQ approaches for ViTs, such as PSAQ V1/V2 (Li et al. 2022b, 2023a), employ a patch similarity metric to separate foreground from background in images. However, these methods overlook the overall image structure and the positional context of patches, potentially resulting in poor quality synthetic images.

To this end, we propose MimiQ, a DFQ framework for low-bit ViT quantization by focusing on inter-head atten-

*Corresponding author

tion similarity. By inspecting the attention maps of real and synthetic data, we observe that synthetic samples show misaligned attention maps, and aligning those maps improves accuracy (Fig. 1). Inspired by this motivational study, we design a DFQ method to achieve inter-head attention similarity. In data generation, we align inter-head attention maps of synthetic samples by minimizing the distance between head-wise maps from each spatial query patch. For fine-tuning, we employ head-wise structural attention distillation on the quantized network to mimic its full-precision counterpart.

We extensively evaluate MimiQ across various tasks, ViT networks, and bit settings. The experimental results show that MimiQ outperforms baselines by a significant margin especially in low-bit settings, reducing the gap between data-free and real-data quantization. As a result, MimiQ sets a new state-of-the-art for the data-free ViT quantization.

Our primary contributions are summarized as follows:

- We discover DFQ baselines produce misaligned attention maps across attention heads, and aligning attention maps contributes to the quantization accuracy.
- We propose a synthetic data generation method to align inter-head attention by reducing the structural distance between attention heads output from each query patch.
- We propose a head-wise attention distillation method aligning the structure of attention outputs of quantized networks with those of full-precision teachers.
- The experiments on various tasks and ViT architectures show that MimiQ achieves new state-of-the-art performance for data-free ViT quantization.

Backgrounds

ViT Architectures and Multi-Head Attention

ViT (Dosovitskiy et al. 2021) is an adaptation of Transformer from NLP (Vaswani et al. 2017) to vision. Each Transformer block comprises a multi-head self-attention (MSA) layer and a feed-forward layer. For the length N_d input sequence with d-dimension, $X_{\in N_d \times d}$, MSA performs attention using multiple heads to obtain diverse features as:

$$MSA(X) = [H_1(X), \cdots, H_N(X)]W^O, \tag{1}$$

where N is the number of attention heads. The outputs of each head are concatenated ($[\cdot]$) and merged by multiplication with projection matrix W^O. Each attention head has separated weights (W_h^Q, W_h^K, W_h^V) for computing query, key, and value vectors. The output of h-th head is as follows:

$$(Q_h, K_h, V_h) = (XW_h^Q, XW_h^K, XW_h^V) \tag{2}$$

$$H_h(X) = softmax(\frac{Q_h K_h^\intercal}{\sqrt{d}})V_h. \tag{3}$$

Data-Free Quantization

Quantization reduces network complexity by converting floating-point to integer operations (Nagel et al. 2021; Gholami et al. 2021). We employ uniform quantization which uses a simple scale (s) and zero-point (z) mapping to transform floating-point values θ into integers θ^{int}:

$$\theta^{int} = clamp(\lfloor \theta \cdot s - z \rceil, q_{min}, q_{max}), \tag{4}$$

where (q_{min}, q_{max}) indicates minimum and maximum of the integer representation range, i.e., (-2^{k-1}, $2^{k-1}-1$). Refer to the extended version for more details and note that this work is not limited to a certain quantization scheme. One drawback of quantization is the potential accuracy loss due to reduced precision. To counter this, quantization-aware training (QAT) uses fine-tuning to regain lost accuracy. However, the training dataset is not always available in real-world scenarios (Liu et al. 2021a; Hathaliya and Tanwar 2020), making QAT inapplicable.

Data-Free Quantization aims to quantize pretrained networks without access to any of the real training data, mostly by using synthetic samples as surrogates. The major challenge is that one cannot use the training data or external generators for the synthesis, as it would fall into a case of data leakage. Instead, information from pretrained full-precision networks f is used by optimizing the following terms:

$$\mathcal{L}_{CL} = -\sum_{c=1}^{C} \hat{y}_c log(f(I)_c), \tag{5}$$

$$\mathcal{L}_{TV} = ||I_{h,w} - I_{h+1,w}||_2^2 + ||I_{h,w} - I_{h,w+1}||_2^2, \tag{6}$$

$$\mathcal{L}_{BNS} = \|\hat{\mu}_l - \mu_l\|_2^2 + \|\hat{\sigma}_l - \sigma_l\|_2^2, \tag{7}$$

where I is the synthetic image, $\hat{y}_c$ is a class label among C classes, (h,w) are pixel coordinates, and (μ_l,σ_l) are BN statistics of the l-th layer. $\mathcal{L}_{CL}$ embeds prior knowledge of the pretrained classifier, and $\mathcal{L}_{TV}$ prevents steep changes between nearby pixels. $\mathcal{L}_{BNS}$ reduces the distance between feature statistics of synthetic samples and BN layers, but it is not applicable to ViTs due to its lack of BN layers.

Related Work

Vision Transformer Quantization

After the success of ViTs, many efforts have been followed to reduce its computational and memory costs through quantization. One of the pioneering efforts is PTQ-ViT (Liu et al. 2021c), which performed quantization to preserve the functionality of the attention. Then followed FQ-ViT (Lin et al. 2022) proposed to fully quantize ViT, including LayerNorm and Softmax. PTQ4ViT (Yuan et al. 2022) applied twin uniform quantization strategy and Hessian-based metric for determining scaling factor. I-ViT (Li and Gu 2023) performed integer-only quantization without any floating-point arithmetic. Q-ViT (Li et al. 2022a) and RepQ-ViT (Li et al. 2023b) proposed remedies to overcome accuracy degradation in low-bit ViTs. However, they require the original training data for calibration, and do not consider real-world scenarios where training data is often unavailable.

Data-Free Vision Transformer Quantization

After the first proposal for data-free quantization (Nagel et al. 2019), many efforts specialized for CNNs have followed, including ZeroQ (Cai et al. 2020), DSG (Zhang et al. 2021), and intraQ (Zhong et al. 2022). Notably, GDFQ (Xu et al. 2020) proposed to jointly train generators to synthesize samples, which laid the foundation for variants using better generators (Zhu et al. 2021), boundary samples (Choi et al. 2021), smooth loss surface (Choi et al. 2022), and sample

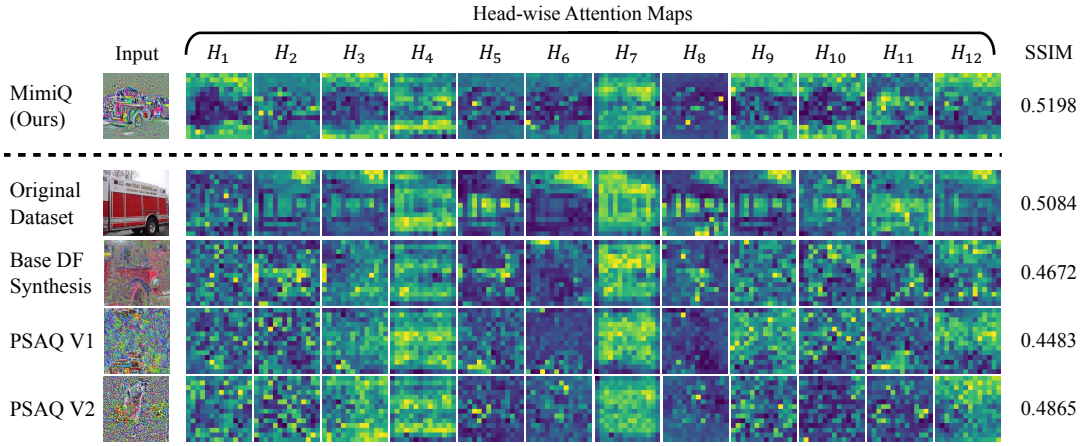


Figure 2: Attention visualization of synthetic samples and the original dataset. Compared to the original dataset, synthetic data from baselines present misaligned attention maps. The first row, a sample generated by MimiQ, shows aligned attention maps across attention heads. The measured average attention similarity (SSIM) further validates MimiQ shows the best alignment.

adaptability (Qian et al. 2023). This stream of methods owes a large portion of its success to the BN statistics (Eq. (7)).

Unfortunately, the BN layer is absent in ViTs, making the CNN-targeted techniques suffer from inaccurate sample distribution when adopted to ViTs. DFQ for ViT is still in an early stage of development. PSAQ-ViT (Li et al. 2022b) was the first to present DFQ for ViTs, utilizing inter-patch discrepancy of foreground and background patches to generate realistic samples. The following work, PSAQ-ViT V2 (Li et al. 2023a), uses an adaptive method by applying the concept of adversarial training. However, they only focus on patch-level similarity, neglecting overall image structure. Parallel to our work, Ramachandran, Kundu, and Krishna (2024), recently disclosed online, explores patch-level contrastive learning to enhance the synthetic data, but still only considers patch-level information similar to prior works. As their method does not consider low-bit regions and has different quantization settings from ours, we believe it is not quite adequate to compare ours with their reported results.

Motivational Study

All the baseline DFQ methods experience huge accuracy drops in low-bit settings compared to the real-data QAT (Tab. 2). To investigate the source of such discrepancy, we inspect attention maps of ViTs, specifically the ViT's head-wise attention maps, denoted as H_i in Eq. (1).

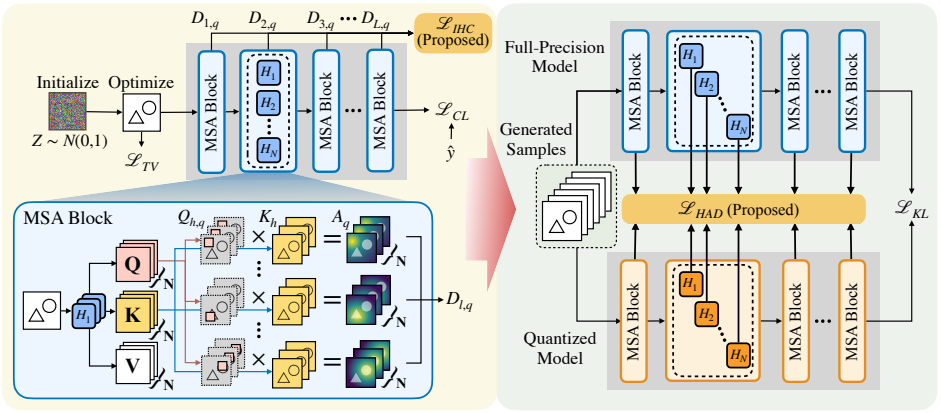
The visualization in Fig. 2 shows that real samples lead to similarly structured attention maps, unlike data-free synthetic samples. We set the base DF synthesis method, which generates synthetic samples with $\mathcal{L}_{CL}$ (Eq. (5)) and $\mathcal{L}_{TV}$ (Eq. (6)), where we also analyze PSAQ V1 and V2 as additional baselines. On the one hand, the attention maps from real images clearly display the object's structure in most heads. While minor variations exist in different heads highlighting different parts, either the object itself (e.g., H_{11}) or the background (e.g., H_9), they exhibit high similarity to one another. On the other hand, synthetic samples from baselines do not seem to produce visually similar features across attention heads. In addition, we present a quantitative analysis with the SSIM score. In the last column of Fig. 2, the baselines show lower attention similarity compared to the real samples. The results also show that MimiQ show the best alignment of attention heads. Fig. 2 shows visualizations from ViT-Base architecture with 12 attention heads. Please refer to the extended version for more results.

From the observation in Fig. 2, we hypothesize that aligning inter-head attention from synthetic samples contributes to better accuracy of data-free quantized ViTs. To validate this, we performed motivational experiments to identify the correlation between attention map similarity and quantization accuracy. First, we generate a synthetic dataset using the base DF synthesis mentioned above. We then measure the inter-head attention similarity of each image with structural similarity index measure (SSIM, Wang et al. (2004)) and construct subsets of the synthetic dataset having 1) high attention similarity and 2) low attention similarity. For comparison, we construct a control group with 3) random sampling. Lastly, we train a W4/A4 quantized ViT network with each sampled group and examine the accuracy.

The results of the experiments (Fig. 1) show that the quantized networks trained with samples of low attention similarity consistently underperform compared to networks trained with samples of high attention similarity. These results empirically validate our hypothesis that the inter-head attention similarity of synthetic samples correlates with the quantization accuracy. Based on the observation, we devise MimiQ to encourage inter-head attention similarity throughout the whole DFQ process, including both data generation and fine-tuning phases. As shown with turquoise bars in Fig. 1, samples generated with MimiQ yield higher attention similarity and have superior quantization accuracy.

Proposed Method

Inspired by the observation from the motivational study, we propose MimiQ framework, a DFQ framework for ViTs that utilizes head-wise similarity information from MSA layers.



(a) Synthetic Data Generation (b) Quantized Network Fine-tuning

Figure 3: An overview of the proposed method. (a) The synthetic samples are initialized with Gaussian noise, then optimized with $\mathcal{L}_G$ (Eq. (10)). The proposed inter-head coherency loss $\mathcal{L}_{IHC}$ measures the similarity of head-wise attention maps from the same query patch index. (b) The quantized network is trained to minimize output ($\mathcal{L}_{KL}$) and inter-head ($\mathcal{L}_{HAD}$) discrepancy.

The overall process is depicted in Fig. 3. We promote inter-head similarity in two directions of sample synthesis and head-wise distillation during fine-tuning.

Sample Synthesis with Inter-Head Similarity

We propose synthetic sample generation with inter-head attention similarity by aligning head-wise internal attention maps. The Fig. 3a depicts attention head alignment. First, we collect attention maps from each head that shares the same spatial query patch index q out of P patches:

$$A_q = [Q_{1,q}K_1^\intercal \quad Q_{2,q}K_2^\intercal \quad \cdots \quad Q_{N,q}K_N^\intercal], \tag{8}$$

where A_q is the collected attention map and (Q,K) are query and key matrices from Eq. (2), respectively. Then, we measure the average distance D_q with attention distance metric f_{dist} between the attention heads as follows:

$$D_q = \frac{1}{N^2}\sum_i^N \sum_j^N f_{dist}(A_{q,i}, A_{q,j}). \tag{9}$$

Here, f_{dist} needs to consider the nature of attention maps: Attention maps can be inverted while retaining the structure of images. For example, in the original sample in Fig. 2, H_{11} focuses on an object while H_9 focuses on the background. The metric should consider such relation among the heads.

To this end, we use absolute SSIM as f_{dist}. A higher magnitude of SSIM indicates a higher correlation in both the positive and negative (inverted) directions, representing the structural similarity between two attention maps. Using Eq. (9), we can optimize synthetic samples towards inter-head attention similarity with synthesis loss $\mathcal{L}_G$.

$$\mathcal{L}_G = \mathcal{L}_{IHC} + \alpha\mathcal{L}_{CL} + \beta\mathcal{L}_{TV}, \tag{10}$$

$$\mathcal{L}_{IHC} = \frac{1}{LP}\sum_l^L \sum_q^P (1 - D_{l,q}), \tag{11}$$

where α and β are hyperparameters for $\mathcal{L}_{CL}$ (Eq. (5)) and $\mathcal{L}_{TV}$ (Eq. (6)), respectively, L is the number of MSA layers in the model, and $D_{l,q}$ is D_q from layer l. The generated synthetic samples can be found in Fig. 5 and Appendix in the extended version.

Corr. Coeff.	DSSIM	MSE	L1-Dist.	KL-Div.
Spearman	**0.9970**	0.9823	0.9796	0.9657
Kendall	**0.9520**	0.8937	0.8863	0.8566

Table 1: Absolute correlation coefficients of attention coherency metric to the quantized accuracy.

Head-Wise Structural Attention Distillation

Here, we propose head-wise structural distillation from a full-precision teacher shown in Fig. 3b, in addition to utilizing our attention-aligned samples. Along with the output matching loss (i.e., $\mathcal{L}_{KL}$) commonly adopted for QAT, we further reduce the distance $g_{dist}(\cdot)$ between each attention output pair by optimizing the following objective:

$$\mathcal{L}_{HAD} = \frac{1}{LN}\sum_l^L \sum_i^N g_{dist}(H_{l,i}^{\mathcal{T}}, H_{l,i}^{\mathcal{S}}), \tag{12}$$

where $\mathcal{L}_{HAD}$ is head-wise attention distillation loss. $H_{l,i}^{\mathcal{T}}$ and $H_{l,i}^{\mathcal{S}}$ are i-th attention head outputs from the l-th layer of teacher and student, respectively. Therefore, the training objective $\mathcal{L}_T$ of the quantized network is as follows:

$$\mathcal{L}_T = \mathcal{L}_{KL}(f_{\mathcal{T}}(\hat{X})||f_{\mathcal{S}}(\hat{X})) + \gamma\mathcal{L}_{HAD}, \tag{13}$$

where $\hat{X}$ is synthetic samples, γ is a hyperparameter.

We compare four candidate metrics for g_{dist}: Mean-squared error (MSE), L1 distance, KL-divergence, and structural dissimilarity (DSSIM), i.e., the negative of SSIM. To choose a metric relevant to quantization accuracy, we randomly quantized a portion of attention heads in each MSA layer of the pretrained ViT-Base network and measure the attention head distance and network accuracy. We sample 500 settings from the configuration space and report Spearman and Kendall rank correlation coefficients. The comparison is shown in Tab. 1. The results show that DSSIM has the highest correlation with quantized accuracy. According to the experimental results, we choose DSSIM as g_{dist}. Please refer to extended version for data visualization.

Bits	Methods	Target Arch.	Networks								
			ViT-T	ViT-S	ViT-B	DeiT-T	DeiT-S	DeiT-B	Swin-T	Swin-S	Swin-B
W4/A4	Real-Data FT	-	58.17	67.21	67.81	57.98	62.15	64.96	73.08	76.34	73.06
	GDFQ	CNN	2.95	4.62	11.73	25.96	22.12	30.04	42.08	41.93	36.04
	Qimera	CNN	0.57	7.02	5.61	15.18	11.37	32.49	47.98	39.64	29.27
	AdaDFQ	CNN	2.00	1.78	6.21	19.57	14.44	19.22	38.88	39.40	32.26
	PSAQ-ViT V1	ViT	0.67	0.15	0.94	19.61	5.90	8.74	22.71	9.26	23.69
	PSAQ-ViT V2	ViT	1.54	4.14	2.83	22.82	32.57	45.81	50.42	39.10	39.26
	MimiQ (Ours)	ViT	**42.99**	**55.69**	**62.91**	**52.03**	**62.72**	**74.10**	**69.33**	**70.46**	**73.49**
	Acc. Gain		+40.04	+48.68	+51.18	+26.07	+30.15	+28.28	+18.91	+28.53	+34.23
W5/A5	Real-Data FT	-	68.49	73.90	80.52	66.10	73.95	78.39	78.71	81.74	83.08
	GDFQ	CNN	24.40	53.96	33.56	44.76	57.00	71.03	61.30	78.04	70.55
	Qimera	CNN	26.70	16.13	9.43	33.13	33.65	47.01	62.13	46.81	43.57
	AdaDFQ	CNN	27.10	59.36	43.02	53.85	59.55	71.12	64.61	79.82	75.59
	PSAQ-ViT V1	ViT	17.66	23.37	16.80	53.36	47.35	57.23	58.63	76.33	57.80
	PSAQ-ViT V2	ViT	40.21	63.59	74.29	55.18	65.30	73.16	69.77	80.55	79.80
	MimiQ (Ours)	ViT	**62.40**	**70.02**	**78.09**	**63.40**	**72.59**	**78.20**	**76.39**	**80.75**	**82.05**
	Acc. Gain		+22.19	+6.43	+3.80	+8.22	+7.28	+5.04	+6.63	+0.20	+2.25
W4/A8	Real-Data FT	-	71.52	79.84	84.52	70.37	78.93	81.47	80.47	82.46	84.29
	GDFQ	CNN	62.65	76.06	81.68	65.82	76.49	80.03	78.90	81.47	83.63
	Qimera	CNN	61.80	60.08	63.22	61.90	70.10	72.38	73.93	72.22	76.35
	AdaDFQ	CNN	64.67	76.27	82.43	67.71	76.92	80.49	79.70	82.07	83.78
	PSAQ-ViT V1	ViT	59.59	62.98	67.74	66.16	76.56	80.05	79.06	81.89	79.51
	PSAQ-ViT V2	ViT	66.78	78.24	84.02	68.23	78.27	81.15	79.98	82.04	83.90
	MimiQ (Ours)	ViT	**68.15**	**78.77**	**84.20**	**69.86**	**78.48**	**81.34**	**80.06**	**82.08**	**83.99**
	Acc. Gain		+1.37	+0.53	+0.18	+1.63	+0.21	+0.20	+0.08	+0.01	+0.09
W8/A8	Real-Data FT	-	74.83	81.30	85.13	71.99	79.70	81.77	80.96	83.08	84.79
	GDFQ	CNN	72.90	80.97	84.81	71.83	79.59	81.62	80.83	82.99	84.42
	Qimera	CNN	72.88	81.04	84.98	71.76	79.46	81.58	80.41	82.95	84.37
	AdaDFQ	CNN	73.84	81.11	84.88	71.72	79.34	81.73	80.89	82.99	84.70
	PSAQ-ViT V1	ViT	72.73	81.17	84.89	71.99	79.71	81.79	**81.26**	**83.29**	**85.13**
	PSAQ-ViT V2	ViT	73.43	81.25	85.11	71.90	79.70	81.86	80.88	83.00	84.71
	MimiQ (Ours)	ViT	**74.60**	**81.30**	**85.17**	**72.01**	**79.73**	**81.87**	80.96	83.07	84.79
	Acc. Gain		+0.76	+0.05	+0.07	+0.02	+0.02	+0.01	-0.30	-0.22	-0.34

Table 2: Comparison on ImageNet image classification dataset.

Performance Evaluation

Experimental Settings

We evaluate MimiQ using tiny, small, and base versions of ViT (Dosovitskiy et al. 2021), DeiT (Touvron et al. 2021), and Swin Transformer (Liu et al. 2021b). We conduct benchmarks on ImageNet classification (Krizhevsky, Sutskever, and Hinton 2012), COCO object detection (Lin et al. 2014), and ADE20K semantic segmentation tasks (Zhou et al. 2019). We used min-max and LSQ (Esser et al. 2020) quantization and reported the best performance. We generated 10k samples with 2k optimization steps per batch with α=1.0, and β=2.5e-5, following Yin et al. (2020). For fine-tuning, we used γ=$\{1.0, 10.0, 100.0\}$, training for 200 epochs. We adapted data augmentations from SimCLR (Chen et al. 2020) for synthetic data generation and training of MimiQ. Please refer to the extended version for the details.

Comparison on Image Classification

The experimental results are presented in Tab. 2. We first provide "Real-Data FT" accuracies from QAT with the original training dataset, which are considered as the empirical upper bound accuracy of DFQs. Then, for the DFQ designed for CNN, we utilize all components applicable to ViTs.

Overall, MimiQ shows significant accuracy gain in low-bit settings and various network, with a maximum gain of 51.18%p. In some cases (DeiT-S/B, Swin-B), MimiQ even outperforms Real-Data FT due to the proposed structural attention head distillation, which provides better guidance to follow full precision attention under a high compression rate. The results from the W4/A8 and W8/A8 settings show that MimiQ achieves similar performance compared to the Real-Data FT without access to any real samples.

In Tab. 2, the quantization accuracies of DeiT are significantly higher than those of similar-size ViTs. This may be due to the stronger inductive bias of DeiT compared to ViT, which enhances robustness against perturbations and preserves its capability under quantization noise.

Object Detection and Semantic Segmentation

The results on the COCO object detection task in Tab. 3 show that MimiQ recovers from quantization error by outperforming the baseline in most settings. The performance gains in low-bit settings are highly noticeable, achieving up to 22.85%p gain. In contrast, baseline methods nearly cause the network to collapse in the W4/A4 setting.

Results on the ADE20K semantic segmentation task also show greate improvements on low-bit settings. The DeiT-S backbone achieves high performance gain, as it suffers from notable degradation compared to the Swin backbones. This is because DeiTs utilize a weak inductive bias compared to Swin, which adapts architectural inductive bias. Therefore,

Bits	Methods	COCO Dataset				ADE20K dataset			
		Swin-T		Swin-S		Backbones (mIoU)			
		AP_{box}	AP_{mask}	AP_{box}	AP_{mask}	DeiT-S	Swin-T	Swin-S	Swin-B
W4 A4	Real FT	31.17	30.75	37.89	36.44	27.47	37.76	44.36	43.28
	PSAQ V1	0.06	0.06	0.05	0.06	0.15	1.65	3.30	0.89
	PSAQ V2	4.52	5.03	12.12	12.20	2.60	3.83	12.13	6.33
	MimiQ	**26.41**	**26.63**	**34.97**	**33.53**	**17.20**	**29.92**	**38.29**	**36.40**
	Gain	+21.89	+21.60	+22.85	+21.33	+14.60	+26.09	+26.16	+30.07
W5 A5	Real FT	42.98	39.66	46.61	42.18	33.10	40.13	47.14	47.43
	PSAQ V1	0.41	0.46	0.64	0.63	0.80	20.26	33.10	39.36
	PSAQ V2	32.69	31.21	45.20	40.99	5.35	26.35	37.58	42.01
	MimiQ	**41.63**	**38.53**	**46.13**	**41.89**	**28.84**	**38.88**	**45.68**	**45.66**
	Gain	+8.94	+7.32	+0.93	+0.90	+23.49	+12.53	+8.10	+3.65
W4 A8	Real FT	39.55	38.00	43.34	41.09	40.96	42.77	47.56	47.63
	PSAQ V1	33.45	32.97	37.57	36.35	35.73	41.25	46.42	46.70
	PSAQ V2	38.71	**37.59**	42.69	40.70	16.92	42.29	46.22	46.65
	MimiQ	**38.77**	37.58	**42.77**	**40.87**	**41.18**	**43.24**	**46.91**	**47.49**
	Gain	+0.06	-0.01	+0.08	+0.17	+5.45	+0.95	+0.49	+0.79
W8 A8	Real FT	46.01	41.63	48.29	43.13	41.96	43.62	46.16	45.39
	PSAQ V1	39.54	36.31	44.20	39.92	38.99	44.36	**47.68**	47.83
	PSAQ V2	45.84	41.51	48.17	43.22	19.51	44.26	47.56	47.68
	MimiQ	**46.03**	**41.58**	**48.31**	**43.25**	**41.76**	**44.39**	47.62	**47.87**
	Gain	+0.19	+0.07	+0.14	+0.03	+2.77	+0.03	-0.06	+0.04

Table 3: Comparison on COCO and ADE20K dataset.

$\mathcal{L}_G$			$\mathcal{L}_T$	Network			
$\mathcal{L}_{IHC}$	$\mathcal{L}_{CL}$	$\mathcal{L}_{TV}$	$\mathcal{L}_{HAD}$	ViT-T	DeiT-T	ViT-B	DeiT-B
✗	✓	✓	✗	13.28	37.70	7.72	34.84
✓	✗	✓	✗	12.31	16.29	0.59	6.78
✓	✓	✗	✗	38.38	50.21	37.80	56.88
✓	✓	✓	✗	39.61	50.16	39.67	62.11
✓	✓	✓	✓	42.99	52.03	62.91	74.10

Table 4: Ablation study of the loss choices.

DeiT backbones are vulnerable to quantization noise, as they need to preserve more information in their parameters.

Analysis

Sensitivity and Ablation Study

We conduct an ablation study of the individual effect of loss functions, shown in Tab. 4. Regarding synthesis loss $\mathcal{L}_G$, we see accuracy drops when $\mathcal{L}_{CL}$ is excluded due to a lack of crucial class information from the pretrained classifier. As $\mathcal{L}_{TV}$ only regularizes steep changes across nearby pixels, it has minor impact on quantization accuracy. Overall, the best results are achieved when all losses are applied in $\mathcal{L}_G$. Also, the proposed distillation loss $\mathcal{L}_{HAD}$ boosts the quantization accuracy by up to 23.24%p. This method is especially effective on larger models (ViT/DeiT-B) due to their higher number of attention heads, allowing for better guidance.

Tab. 5 shows the sensitivity of each hyperparameter α, β, and γ (Eq. (10), Eq. (13)), where experiment is conducted by varying each hyperparameter while others are fixed to the default value. The base networks perform better at higher α, indicating that larger models can embed more class-related information in samples. Also, models with more heads favor higher head distillation factor γ. In contrast, β is network-insensitive, as it only considers changes in the pixel value.

α	0.01	0.1	1	10
ViT-T	41.41	42.54	42.67	41.04
ViT-B	37.09	49.87	53.96	55.32
DeiT-T	41.22	49.58	52.03	50.96
DeiT-B	54.07	60.02	63.11	63.52
β	2.5E-07	2.5E-06	2.5E-05	2.5E-04
ViT-T	41.17	41.19	42.67	41.56
ViT-B	48.61	48.61	53.96	51.27
DeiT-T	50.90	50.85	52.03	48.36
DeiT-B	63.77	63.13	63.11	63.13
γ	0.1	1	10	100
ViT-T	40.05	42.67	42.99	36.48
ViT-B	51.36	53.96	62.34	62.91
DeiT-T	50.65	52.03	51.01	44.34
DeiT-B	62.12	63.11	70.23	74.10

Table 5: Sensitivity analysis of hyperparameters.

Network	Generation (g_{dist})				Distillation (f_{dist})			
	MSE	L1 Dist.	KL Div.	SSIM	MSE	L1 Dist.	KL Div.	SSIM
ViT-T	40.91	40.57	39.63	**42.99**	4.71	37.29	40.38	**42.99**
DeiT-T	50.88	50.60	50.33	**52.03**	17.77	49.33	49.17	**52.03**
ViT-B	44.90	44.56	40.96	**62.91**	23.47	53.25	59.23	**62.91**
DeiT-B	63.83	63.97	62.59	**74.10**	68.04	69.78	69.59	**74.10**

Table 6: Performance comparison on coherency metrics.

Inter-Head Attention Similarity Metrics

We provide a comparison of attention similarity metrics on sample synthesis and attention distillation. Tab. 6 shows that SSIM-based synthesis exhibits superior performance, while L1 distance and KL-divergence show lesser effectiveness. For attention-head distillation, Tab. 6 presents similar trends that SSIM-based distillation achieves the highest accuracy, which agrees with the rank correlation coefficients in Tab. 1.

Computational Costs for Quantization

The correlation between synthetic dataset size and quantization accuracy (Fig. 4) reveals that MimiQ performs robustly with varying dataset sizes, surpassing baselines with only 64 samples in W4/A4 settings and demonstrating reasonable performance with 1k samples. We then compare computational costs in Tab. 7. In addition to the default setting of MimiQ using 10k samples, we also present results with similar computational costs to the baselines, where -nk indicates the size of the training dataset. The comparison indicates MimiQ outperforms baselines with similar or fewer costs, demonstrating the cost efficiency of MimiQ.

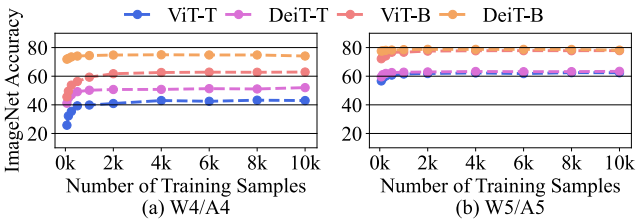
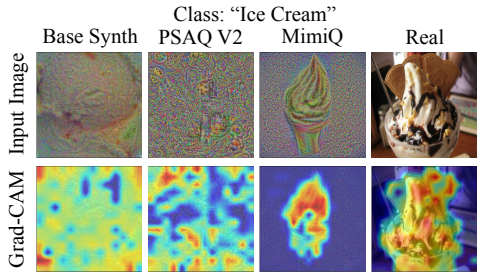


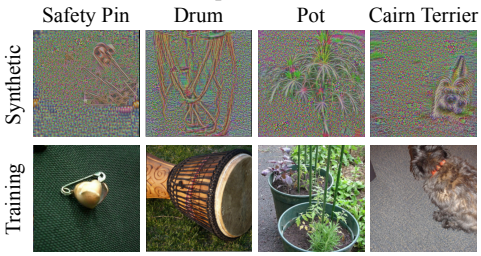
Figure 4: Sensitivity analysis of synthetic dataset size.

Method	Type	Synth.	Quant.	Total	Acc.
GDFQ	QAT	-	10.70h	10.70h	11.73
AdaDFQ	QAT	-	8.44h	8.44h	6.21
PSAQ V1	PTQ	0.11h	0.0002h	0.11h	0.94
PSAQ V2	QAT	-	4.55h	4.55h	2.83
MimiQ-1k	QAT	1.98h	2.39h	4.37h	**59.32**
MimiQ-4k	QAT	7.92h	2.39h	10.31h	**62.59**
MimiQ-10k	QAT	19.79h	2.39h	22.18h	**62.91**

Table 7: Quantization cost comparison of ViT-B network.



(a) Grad-CAM comparison of DFQ methods.



(b) Pairs of real/synthetic images with the lowest LPIPS. There is no indication of a privacy leak.

Figure 5: Grad-CAM and LPIPS analysis of MimiQ.

Grad-CAM Analysis

As part of our analysis of how the MimiQ-generated images are viewed from the network's perspective, we utilize Grad-CAM (Selvaraju et al. 2017) to visualize the attention map from the last layer (Fig. 5a). It can be observed that MimiQ-generated images have most of their object information clear and well-clustered, similar to the real images. On the other hand, images from other methods have much of their object information scattered, which could harm the accuracy.

Discussion

Adaptation to Real-Data ViT Quantization. To further explore the efficacy of MimiQ, we benchmark adaptation of MimiQ samples to real-data ViT quantization methods (PTQ4ViT, FQ-ViT, RepQ-ViT) following the original settings of each paper. The results are shown in Tab. 8, where "Attn$k$" refers to k-bit attention quantization. Please note that these experiments utilize only synthetic samples from the MimiQ framework, replacing the proposed attention distillation phase with existing quantization methods. In most cases, calibration with MimiQ samples achieves an accuracy similar to that of real-data calibration. Notably, our results

Methods	Bits	Calib. data	Models		
			ViT-B	DeiT-B	Swin-B
PTQ4ViT	W6/A6	Real data	81.65	80.25	84.01
		Synth. (MimiQ)	82.40	80.43	83.96
	W8/A8	Real data	84.25	81.48	85.14
		Synth. (MimiQ)	84.66	81.49	85.25
FQ-ViT	W8/A8/Attn4	Real data	82.68	80.85	82.38
		Synth. (MimiQ)	81.78	80.75	82.11
	W8/A8/Attn8	Real data	83.31	81.20	82.97
		Synth. (MimiQ)	82.52	81.19	82.89
RepQ-ViT	W4/A4	Real data	68.48	75.61	78.32
		Synth. (MimiQ)	19.96	72.15	67.00
	W6/A6	Real data	83.62	81.27	84.57
		Synth. (MimiQ)	82.35	80.84	82.36

Table 8: Comparison with real-data ViT quantizations.

ImageNet Acc. (%)	Synthetic/Real Distinguishability	Synthetic→Real Transferability
Train	99.97	49.69
Test	99.99	0.16

Table 9: Experiments on model inversion and identity attacks using ResNet-18. Results indicate the attacks fail.

on PTQ4ViT show that MimiQ often outperforms the original results, suggesting MimiQ's potential to enhance real-data quantization. However, we observe accuracy drops in RepQ-ViT in the W4/A4 setting. This may be due to the overfitting to synthetic samples leading to distribution shifts, where MimiQ framework counters this by minimizing head-wise attention discrepancy. Overall, the results demonstrate MimiQ's versatility and its potential for broader application.

Does MimiQ Threaten Privacy? MimiQ may be linked to input reconstruction attacks (Oh and Lee 2019; Vepakomma et al. 2021) that resemble specific images of original data. Comparison using LPIPS against the original training dataset reveals that only general features are captured without replicating specific images (Fig. 5b). Further tests following Prakash et al. (Tab. 9) shows that MimiQ samples are distinguishable from real ones with near-perfect train and test accuracies of 99.97% and 99.99%, mitigating identity attack concerns. Finally, a synthetic-trained network from scratch underperformed dramatically (0.16%), suggesting a low risk of model inversion attacks.

Conclusion

In this paper, we propose MimiQ, a DFQ for ViTs inspired by attention similarity. We observe head-wise attention maps of synthetic samples are not aligned and aligning them contributes to the quantization accuracy. From the findings, MimiQ utilizes inter-head attention similarity to better leverage the knowledge instilled in the attention architecture, synthesizing training data that better aligns inter-head attention. In addition, MimiQ utilizes fine-grained head-wise attention map distillation. As a result, MimiQ brings significant performance gain, setting a new state-of-the-art results.

Acknowledgments

This work was partially supported by National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.2022R1C1C1011307), Institute of Information & communications Technology Planning & Evaluation (IITP) (RS-2021II211343 (SNU AI), RS-2022-II220871, RS-2021-II212068 (AI Innov. Hub)), and an unrestricted gift from Google. Disclaimer: Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of Google.

References

Cai, Y.; Yao, Z.; Dong, Z.; Gholami, A.; Mahoney, M. W.; and Keutzer, K. 2020. ZeroQ: A novel zero shot quantization framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*.

Choi, K.; Hong, D.; Park, N.; Kim, Y.; and Lee, J. 2021. Qimera: Data-free quantization with synthetic boundary supporting samples. In *Advances in Neural Information Processing Systems*.

Choi, K.; Lee, H. Y.; Hong, D.; Yu, J.; Park, N.; Kim, Y.; and Lee, J. 2022. It's all in the teacher: Zero-shot quantization brought closer to the teacher. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Courbariaux, M.; Bengio, Y.; and David, J.-P. 2015. BinaryConnect: Training deep neural networks with binary weights during propagations. In *Advances in Neural Information Processing Systems*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Esser, S. K.; McKinstry, J. L.; Bablani, D.; Appuswamy, R.; and Modha, D. S. 2020. Learned step size quantization. In *International Conference for Learning Representations*.

Gholami, A.; Kim, S.; Dong, Z.; Yao, Z.; Mahoney, M. W.; and Keutzer, K. 2021. A survey of quantization methods for efficient neural network inference. *arXiv preprint arXiv:2103.13630*.

Hathaliya, J. J.; and Tanwar, S. 2020. An exhaustive survey on security and privacy issues in Healthcare 4.0. *Computer Communications*.

Hubara, I.; Courbariaux, M.; Soudry, D.; El-Yaniv, R.; and Bengio, Y. 2017. Quantized neural networks: Training neural networks with low precision weights and activations. *The Journal of Machine Learning Research*.

Kong, Z.; Dong, P.; Ma, X.; Meng, X.; Niu, W.; Sun, M.; Shen, X.; Yuan, G.; Ren, B.; Tang, H.; et al. 2022. SPViT: Enabling Faster Vision Transformers via Latency-Aware Soft Token Pruning. In *European Conference on Computer Vision*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*.

Li, Y.; Xu, S.; Zhang, B.; Cao, X.; Gao, P.; and Guo, G. 2022a. Q-ViT: Accurate and Fully Quantized Low-bit Vision Transformer. In *Advances in Neural Information Processing Systems*.

Li, Z.; Chen, M.; Xiao, J.; and Gu, Q. 2023a. PSAQ-ViT V2: Toward Accurate and General Data-Free Quantization for Vision Transformers. *IEEE Transactions on Neural Networks and Learning Systems*.

Li, Z.; and Gu, Q. 2023. I-ViT: integer-only quantization for efficient vision transformer inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Li, Z.; Ma, L.; Chen, M.; Xiao, J.; and Gu, Q. 2022b. Patch Similarity Aware Data-Free Quantization for Vision Transformers. In *Proceedings of the European Conference on Computer Vision*.

Li, Z.; Xiao, J.; Yang, L.; and Gu, Q. 2023b. Repq-vit: Scale reparameterization for post-training quantization of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*.

Lin, Y.; Zhang, T.; Sun, P.; Li, Z.; and Zhou, S. 2022. FQ-ViT: Post-Training Quantization for Fully Quantized Vision Transformer. In *International Joint Conference on Artificial Intelligence*.

Liu, B.; Ding, M.; Shaham, S.; Rahayu, W.; Farokhi, F.; and Lin, Z. 2021a. When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Liu, Z.; Wang, Y.; Han, K.; Zhang, W.; Ma, S.; and Gao, W. 2021c. Post-training quantization for vision transformer.

Nagel, M.; Baalen, M. v.; Blankevoort, T.; and Welling, M. 2019. Data-free quantization through weight equalization and bias correction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Nagel, M.; Fournarakis, M.; Amjad, R. A.; Bondarenko, Y.; van Baalen, M.; and Blankevoort, T. 2021. A white paper on neural network quantization. *arXiv preprint arXiv:2106.08295*.

Oh, H.; and Lee, Y. 2019. Exploring image reconstruction attack in deep learning computation offloading. In *The 3rd International Workshop on Deep Learning for Mobile Systems and Applications*.

Prakash, P.; Ding, J.; Li, H.; Errapotu, S. M.; Pei, Q.; and Pan, M. 2020. Privacy preserving facial recognition against model inversion attacks. In *IEEE GLOBECOM*.

Qian, B.; Wang, Y.; Hong, R.; and Wang, M. 2023. Adaptive Data-Free Quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Ramachandran, A.; Kundu, S.; and Krishna, T. 2024. CLAMP-ViT: Contrastive Data-Free Learning for Adaptive Post-Training Quantization of ViTs. *arXiv preprint arXiv:2407.05266*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision*.

Sung, W.; Shin, S.; and Hwang, K. 2015. Resiliency of deep neural networks under quantization. *arXiv preprint arXiv:1511.06488*.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*.

Vepakomma, P.; Singh, A.; Zhang, E.; Gupta, O.; and Raskar, R. 2021. NoPeek-Infer: Preventing face reconstruction attacks in distributed inference after on-premise training. In *IEEE International Conference on Automatic Face and Gesture Recognition*.

Wang, Z.; Bovik, A.; Sheikh, H.; and Simoncelli, E. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*.

Xu, S.; Li, H.; Zhuang, B.; Liu, J.; Cao, J.; Liang, C.; and Tan, M. 2020. Generative low-bitwidth data free quantization. In *European Conference on Computer Vision*.

Yin, H.; Molchanov, P.; Alvarez, J. M.; Li, Z.; Mallya, A.; Hoiem, D.; Jha, N. K.; and Kautz, J. 2020. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yu, H.; and Wu, J. 2023. A unified pruning framework for vision transformers. *Science China Information Sciences*.

Yuan, Z.; Xue, C.; Chen, Y.; Wu, Q.; and Sun, G. 2022. Ptq4vit: Post-training quantization for vision transformers with twin uniform quantization. In *European Conference on Computer Vision*.

Zhang, X.; Qin, H.; Ding, Y.; Gong, R.; Yan, Q.; Tao, R.; Li, Y.; Yu, F.; and Liu, X. 2021. Diversifying Sample Generation for Accurate Data-Free Quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhong, Y.; Lin, M.; Nan, G.; Liu, J.; Zhang, B.; Tian, Y.; and Ji, R. 2022. IntraQ: Learning Synthetic Images with Intra-Class Heterogeneity for Zero-Shot Network Quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhou, B.; Zhao, H.; Puig, X.; Xiao, T.; Fidler, S.; Barriuso, A.; and Torralba, A. 2019. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*.

Zhu, B.; Hofstee, P.; Peltenburg, J.; Lee, J.; and Alars, Z. 2021. Autorecon: Neural architecture search-based reconstruction for data-free compression. In *International Joint Conference on Artificial Intelligence*.