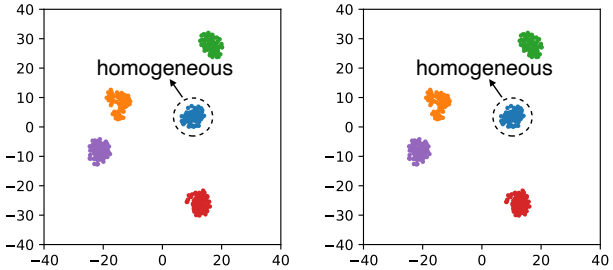
IntraQ: Learning Synthetic Images with Intra-Class Heterogeneity for Zero-Shot Network Quantization (Supplementary Material)

1. MDC *vs*. ADI

In this section, we provide the comparison between the fake data generated by our MDC and ADI [3]. Though both MDC and ADI [3] aim to diversify fake images, their manners are quite different: our MDC manipulates the distances among features of fake images, while ADI enlarges disagreement between the student model and the teacher model. Fig. 1 shows feature visualization of ADI and our MDC: the features of MDC scatter a lot while ADI is in a dense concentration. Besides, using 5,120 synthetic images, our MDC obtains 63.77% top-1 accuracy with ResNet-18 on ImageNet, while ADI only has 54.97% (we use the official code of ADI). Thus, our MDC can produce more diverse synthetic images as well as better performance.



(a) ADI (b) MDC

Figure 1. Feature visualization of ADI and MDC.

2. Other Preprocessing in LOR

In this section, we report the results of using other preprocessing operations. Tab. 1 shows no increase from flip and rotation. This is because our LOR aims to capture informative content at different scales and positions of the synthetic images. Flip and rotation may not benefit this goal.

Operations	L	L+F	L+R	L+R+F
Acc (%)	66.47	66.34	66.31	66.42

Table 1. Results of 4-bit ResNet-18 on ImageNet when adding flip and rotation to our LOR. "L": LOR; "F": flip; "R": rotation.

3. Data Amount

Tab. 2 shows the ablation on amount of synthetic images. Note that we achieve 65.87% accuracy using only 256 images, better than all previous methods, such as the SOTA GZNQ with 64.50% using 100,000 images.

Amount	256	1,280	5,120	10,000	20,000
Acc (%)	65.87	66.14	66.47	66.49	66.50

Table 2. Performance of our IntraQ *w.r.t.* different amounts of synthetic images (4-bit ResNet-18 on ImageNet).

4. More Comparisons

Tab. 3 shows more comparisons with recent ZSQ methods including Qimera [1] and SQuant [2]. We report the top-1 accuracy of 4-bit ResNet on ImageNet. Note that SQuant sets the input of the last layer to 8-bit while our IntraQ and Qimera quantize all layers to 4-bit.

Bit-width	Method	Generator	Acc. (%)
W4A4	Real data	-	67.89
	Qimera	✓	63.84
	SQuant	✗	66.14
	IntraQ (Ours)	✗	**66.47**

Table 3. Results of ResNet-18 on ImageNet. WBAB indicates the weights and activations are quantized to B-bit.

References

[1] Kanghyun Choi, Deokki Hong, Noseong Park, Youngsok Kim, and Jinho Lee. Qimera: Data-free quantization with synthetic boundary supporting samples. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1

[2] 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 *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. 1

[3] Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8715–8724, 2020. 1