Appendices

A. Hyper-Parameters

A.1. Generation

Using ADAM optimizer with learning rate 0.1, betas(0.9,0.999) for all experiments, when image-prior scale is non-zero we use a $\mathcal{N}(0,1)$ smoothing kernel of size 5. Generally batch size is chosen to fit in available memory (statistics are accumulated across devices).

scheme	*batch size*	*in-batch*	*steps*	*lr drops (1e-1)*	*scales (bns,cls,prior)*
ResNet-44-CIFAR10					
BNS	800	4	1000	800	1,0,0
$BNS + \mathcal{I}$					1,1e-3,1
$\mathcal{I}$					0,1e-3,1
Wide ResNet-28-10-CIFAR100					
BNS	256	4	1000	800	1,0,0
$BNS + \mathcal{I}$					1,1e-3,1
$\mathcal{I}$					0,1e-3,1
ResNet-18-ImageNet					
BNS	80	4	1000	800	1,0,0
$BNS + \mathcal{I}$					1,1e-3,1
$\mathcal{I}$					0,1e-3,1
MobileNet-V2-ImageNet					
BNS	400	2	6000	1500,4200	1,0,0
$BNS + \mathcal{I}$			10000	1500,6000	1,1e-4,0
$\mathcal{I}$			1000	800	0,1e-4,1
DenseNet-121-ImageNet					
BNS	200	2	6000	1500,4200	1,0,0
$BNS + \mathcal{I}$			10000	1500,6000	1,1e-4,0
$\mathcal{I}$			1000	800	0,1e-4,1

A.2. Hyper-Parameters sensitivity for calibration

Here we present evidence of hyper-parameter sensitivity for Inception related generation schemes. Specifically, we demonstrate how changing the variance parameter of the Gaussian smoothing kernel, used as the image prior, to regularize the optimization process, impacts the usability of the generated data. In table-5, we compare calibration results for quantized ResNet44 from *Small scale results* section.

Table 5: Post calibration validation accuracy, additional fixed hyper-parameters: kernel 5x5, inception loss scale 0.001. Results on $\mathcal{I}$ and $BNS + \mathcal{I}$ datasets appear to be sensitive to hyper-parameter adjustments. We report our best results without performing an exhaustive search.

sigma	$\mathcal{I}$	$BNS + \mathcal{I}$
ResNet-44, 4w4a, real data calibration accuracy - 89.19 (0.15)		
0.375	89.22 (0.22)	88.87 (0.22)
1.0	87.44 (0.13)	89.1 (0.09)
ResNet-44, 4w4a[1] real data calibration accuracy - 91.64 (0.13)		
0.375	91.26 (0.16)	91.3 (0.1)
1.0	90.89 (0.11)	91.76 (0.22)
ResNet-44, 4w8a[1], real data calibration accuracy - 92.18 (0.05)		
0.375	92.04 (0.03)	92.33 (0.04)
1.0	92.37 (0.03)	92.25 (0.04)

[1]First & final layers are in 8 bits

A.3. Distillation

Algorithm-2 describes the general KD framework with IQ and input-mixup tweaks (see ablation in table-1). These are not strictly required when the full dataset is available, yet tend to offer improved convergence when the available dataset is limited in size. We use

Table 6: Comparison of dataset size and fine-tune objective impact on ImageNet validation accuracy for ResNet-18 quantized to 4 bits. * Regime follows [20].

dataset	*samples*	*objective*	*steps*	*top1*	$\mathcal{J}_{KL}$
ImageNet	1.3M	CE	*550440	69.95	0.05
	1.3M	CE	*45036	68.36	
	100K	CE	44000	67.47	
ImageNet	1.3M	KLD	550440	68.87	0.05
	100K	KLD	44000	68.68	
BNS	100K	KLD	44000	68.14	0.052
$BNS + \mathcal{I}$	100K	KLD	44000	67.95	0.059
BNS+$BNS + \mathcal{I}$	50K+50K	KLD	44000	68.07	0.053

standard KL divergence as KD loss, smoothed-l1 is used for IQ term and the calibration algorithm is as described in section-4.2. We use SGD with a learning rate of 0.1 and a batch size of 256 in all experiments, while sampling with replacement for a fixed number of steps per epoch (CIFAR-200 ,ImageNet-400), additional shared hyper-parameters: $\alpha, \beta, \theta_{mixup}$ = 1, 0.01, {mix_rate=0.5}.

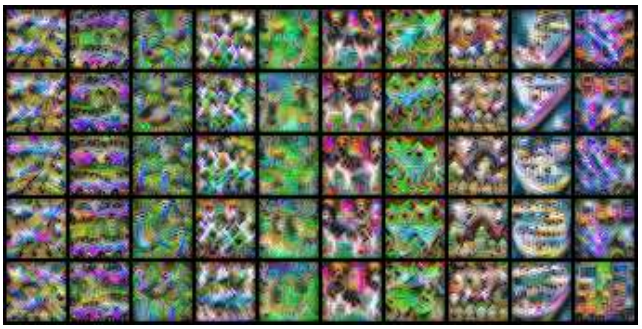
B. Additional results for ImageNet dataset

There exists an accuracy gap between the fine-tuned model and the distilled variants even when the full data-set is available as seen in table-6. We believe this can be attributed to using KD without an additional ground truth loss term which is common in an unsupervised setting. Thus, final accuracy depends solely on the prediction quality of the teacher. Whereas additional label information can be used to penalize the student when repeating similar mistakes made by the teacher and contribute to improved final accuracy. Additionally, we speculate that a given bias in the reference model's prediction towards certain classes (see figure-7) may degrade the student accuracy when training on raw teacher outputs.

C. Generating CIFAR10 and ImageNet

In figure-4, we provide a several synthetic samples per class generated from a CIFAR10 trained ResNet44 model, using Inception scheme ($\mathcal{I}$) and samples generated using the BN-stats + Inception scheme ($BNS + \mathcal{I}$). The generation process follows the settings detailed in *Generating data samples* section. Samples seem to share similar visual features between generation schemes. while $BNS + \mathcal{I}$ samples are smoother and appear clearer.

(a) samples generated with $BNS + \mathcal{I}$, Gaussian smoothing Kernel 5x5, sigma 1.0



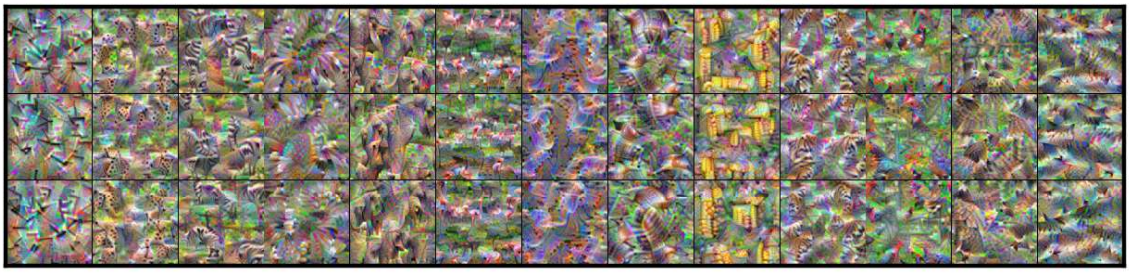
(b) samples generated with Inception loss + Gaussian smoothing Kernel 5x5, sigma 1.0

Figure 4: CIFAR10 samples generated from ResNet-44

Additionally, a subset of samples generated using ResNet-18 model trained over ImageNet dataset is presented in figure-6. Visual inspection of those generated samples indicated considerable improvement in detail and diversity over naive inception generation scheme. Furthermore, some instances seem to preserve the represented class physical structure better than others. Despite the lack of consensus regarding objective image quality evaluation methods, we find it is intriguing to further explore means to improve reproduced class visual quality and investigate its connection to the model's prediction quality. We believe the $BNS + \mathcal{I}$ method may serve as a powerful tool for DNNs interpretability, providing insight into the featured learned by DNNs.

Figure 5: Comparison of ImageNet and synthetic samples generated from ResNet-18

(a) Reference ImageNet samples



(b) $\mathcal{I}$, smoothing kernel 9x9, sigma 0.7 yielded best visual result

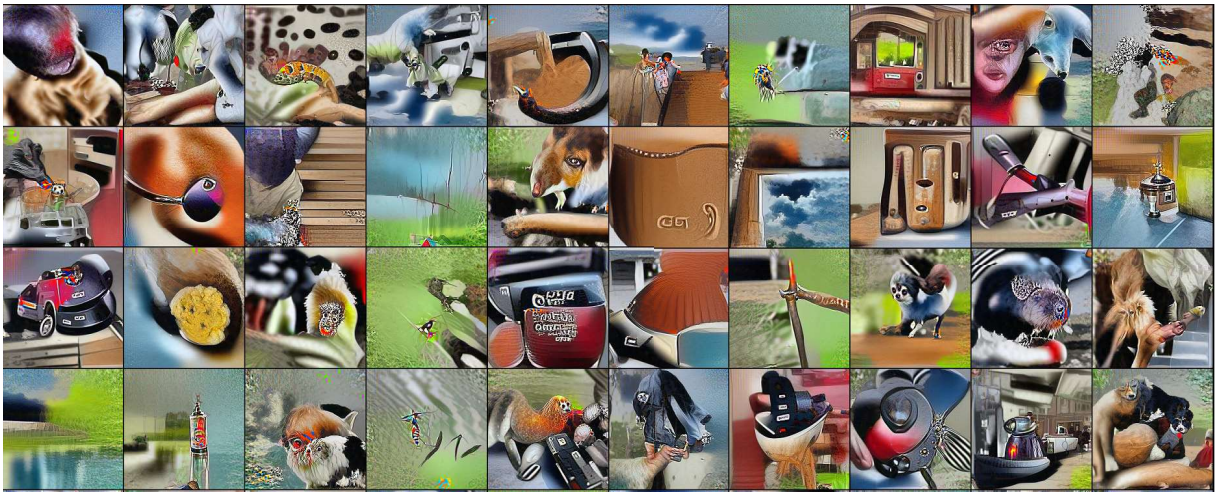
(c) $BNS + \mathcal{I}$, smoothing kernel 5x5, sigma 1

(d) BNS only

D. BN-Stats measurement

In table-7, we provide raw measurements of $\mathcal{J}_{KL}$ on select datasets using pre-trained ResNet44 from table-4 of the main paper.

Figure 6: Additional synthetic samples generated from DenseNet-121



(a) BNS only samples

(b) $BNS + \mathcal{I}$, no image prior used

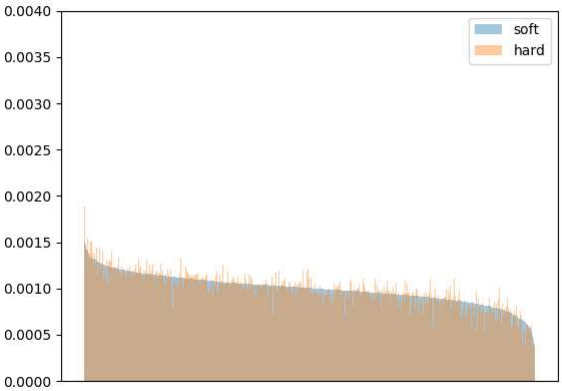
Table 7: $\mathcal{J}_{KL}$ values are computed on the entire training data-set and include all BN layers within the reference model. *Fast Gradient Sign Method (FGSM) was used to create adversarial samples with small perturbation ratio $\epsilon = 0.1$, the measured loss generally grows with epsilon.

Train/Measure	CIFAR10	CIFAR100	MNIST	SVHN	STL10	Random	*FGSM
CIFAR10	0.023	0.022	0.151	0.182	0.031	0.498	0.075
CIFAR100	0.065	0.057	0.223	0.236	0.072	0.845	0.106
MNIST	0.207	0.223	0.002	0.392	0.277	0.655	0.004
SVHN	0.037	0.032	0.032	0.027	0.073	0.111	0.024

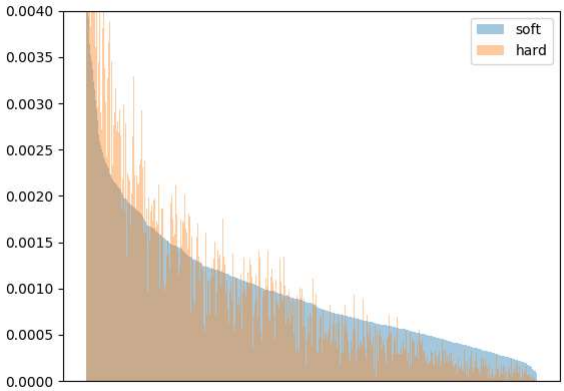
E. Perceived dataset bias

Given a reference classification model, we loosely consider the per-class mean prediction of the model as unbiased if it is close to uniform when presented with a balanced set of examples, since no particular class should be favoured over other classes by an unbiased model. To provide a qualitative evaluation, we record the softmax output of a pretrained ResNet-18 model from *torchvision*, over the entire ImageNet validation set. Then, compute the mean prediction of the "soft" outputs (i.e., $\frac{1}{N}\sum_{i=1}^{N} output_i$) where $output_i \in R^{\#classes}$ is the model softmax output for sample i, as well as the "hard" mean prediction (i.e., $\frac{1}{N}\sum_{i=1}^{N} e^{argmax\{output_i\}}$ where e is the standard

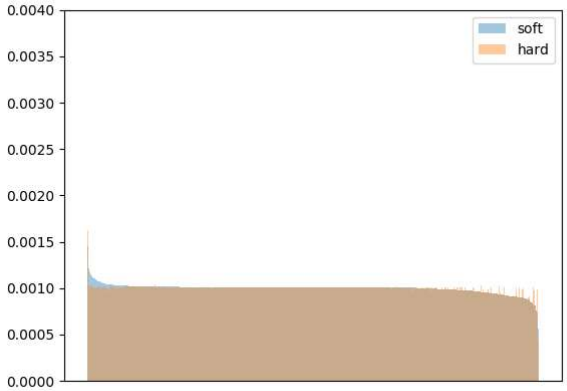
vector). In figures-7a,7b,7c, we plot the mean soft and hard predictions of the model for a given dataset, while classes are sorted according to the mean soft prediction. Figure-7a reveals that although the validation dataset is balanced (ignoring possible annotation errors and class similarity), the model produces a somewhat biased prediction. We hypothesize this bias may impede KD performance as discussed in *Large scale experiment* section, even when the entire dataset is available.



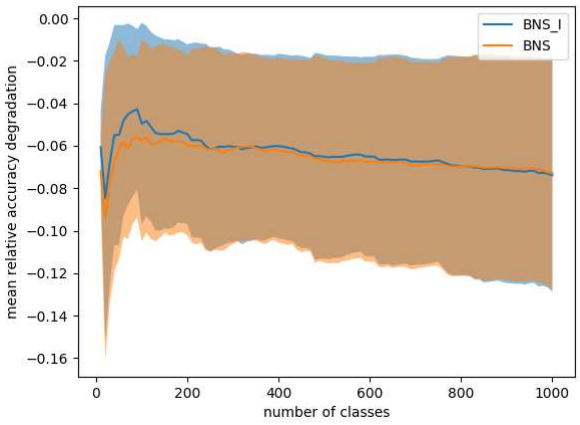
(a) Reference model (fp32) mean prediction over ImageNet validation set reveals a clear classifier bias.



(b) Reference model (fp32) mean prediction over BNS samples presents a highly biased preference towards certain classes.



(c) Reference model (fp32) mean prediction over $BNS + \mathcal{I}$ samples. As expected prediction mean is balanced since the samples are tailored through optimization to favor a single class with uniform target class sampling.



(d) Mean relative tail degradation on ImageNet validation set - classes are sorted according to the hard prediction rate from 7b. Ratio is computed as the mean of $\frac{s^i_{fp32} - s^i_{ft}}{s^i_{fp32}}$ over increasing number of classes from least favored to most favored. s denotes per-class accuracy and i is the class index. The $BNS + \mathcal{I}$ fine-tuned model present a small improvement over the BNS variation with respect to the worst case accuracy degradation.

Figure 7: Bias analysis on ResNet-18, soft and hard plots refer to means of raw model outputs and a hard class index choice (i.e., a hot-1 vector) over the entire validation set

Additionally, in figures-7b,7c we provide the mean prediction plots for BNS and $BNS + \mathcal{I}$ datasets appropriately, each containing 100K synthetic samples. Figure-7b shows a clear preference towards certain classes, which does not necessarily align with the model's prediction bias over the validation set (figure-7a). While figure-7c shows that $BNS + \mathcal{I}$ dataset appears balanced in terms of the mean prediction, which is not a surprising result. However, our KD experiments did not show any significant benefit to using the $BNS + \mathcal{I}$

dataset compared to the BNS dataset in terms of the final validation accuracy, despite the perceived class imbalance.

F. Mean tail degradation

To further investigate the impact of BNS dataset bias on fine-tuned (quantized) model's accuracy, we consider the mean relative accuracy degradation over N least frequent classes (mean tail degradation). First, we measure the per-class validation accuracy for each of the fine-tuned models (i.e., ResNet18 fine-tuned with BNS or $BNS + \mathcal{I}$ dataset, see *Large scale experiment* section). Then, we compute the relative degradation compared to the float model for each class. Finally, classes are sorted according to the reference model (fp32) mean hard prediction over BNS dataset (figure-7b). Essentially, the least frequently predicted classes are considered first under the tail limit (N). Figure-7d shows the mean degradation with an increasing number of classes N, while only considering classes where accuracy degradation is observed (i.e., we ignore classes which present an improvement in accuracy). Figure-7d presents evidence for an improved mean tail degradation, which can serve as motivation for using $BNS + \mathcal{I}$ over BNS dataset since the worst-case accuracy is improved despite the overall accuracy being slightly worse.