This comes as supplementary material to the paper *Rethinking Data Distillation: Do Not Overlook Calibration*. The appendix is structured as follows:

A. Distillation Backbones.

B. Additional Experiments.

A. Distillation Backbones

A.1. Datasets and Networks

Following [11, 10, 1, 2], we use a ConvNet with 3 blocks for CIFAR10 and CIFAR100 [5], a ConvNet with 4 blocks for Tiny-ImageNet [6], and a ConvNet with 5 blocks for Nette (a subset of ImageNet) [4]. Each block in the ConvNets contains a 3×3 convolutional layer with 128 channels, followed by instance normalization [9], ReLU [7] and a 2×2 average pooling layer with stride 2. We apply Kornia ZCA [8] on CIFAR10 and CIFAR100 for distillation backbones [11, 10, 1]. We pick the ConvNet in each distillation backbone because it gives the best distillation performance while keeping the distillation process under an acceptable time and computational budget.

B. Additional Experiments

B.1. Details in Masked Temperature Scaling

We sample from all the distilled data we have as the validation set to update the temperature parameter T in our proposed Masked Temperature Scaling. Instead of sampling from all the shuffled data at once, we perform a per-class sampling such that there is no missing class or over-sampled class, which is especially important for distillation settings that aim for aggressive compression rates such as image-per-class ≤ 10. The traditional temperature scaling [3] separates all the data available into a training set and a validation set and uses the validation set only for updating T. This separated use of the distilled data is not applicable when image-per-class = 1. Moreover, a data split of 10% can hurt training accuracy by as much as 1.68% on the Nette subset of ImageNet, while our proposed during-training calibration method (MDT) only hurts accuracy by 0.24%, as reported in Table 1. In addition, our proposed after-training method Masked Temperature Scaling keeps original training accuracy and achieves better calibration results than temperature scaling as reported in our main text.

B.2. More Results on SVD of Distilled Data and Full Data

As we discussed in our main text, distilled data contain more concentrated information that easily gets grouped by algorithms such as SVD. We here illustrate the cumulative explained ratio of top singular values of data distilled by different backbones. We expect that concentrated information leads to a curve skewed to the top left and evenly distributed information leads to a smooth curve close to the diagonal. This will show how much each component corresponding to the singular values in Σ contributes to the data reconstruction. As shown in Figure 1, the cumulative explained ratio given by ours grows at the most steady rate, showing that our method produces more evenly distributed information in distilled data compared to the overly condensed information in other distillation backbones. As we concluded in our main text, this serves as a regularization to the distillation process such that it cannot discard too much information that is unrelated to the classification task but semantically meaningful for other tasks, leading to more calibratable networks trained on the resulting distilled data.

Table 1. Accuracy (%) drops by as much as 1.68% when training with 90% of distilled Nette (a subset of ImageNet). The rest 10% is used in temperature scaling (TS). Our proposed after-training MTS (shadow) keeps the original accuracy. Our proposed during-training MDT (shadow) keeps a higher accuracy than that of dropping 10% of training data for TS. We use MTT [1] as the distillation backbone.

Dataset	Full, MTS (Ours)	TS (10%)	MDT (Ours)
CIFAR10	70.48 ± 0.2	69.78 ± 0.5	69.98 ± 0.4
CIFAR100	47.47 ± 0.2	47.10 ± 0.2	46.21 ± 0.4
Tiny ImageNet	27.76 ± 0.2	27.35 ± 0.2	27.62 ± 0.4
ImageNette	63.04 ± 1.3	61.36 ± 1.6	62.80 ± 1.2

B.3. More Results on DDNNs' Limited Encoding Capacity

We provide more visualizations of projections of intermediate feature vectors obtained from DDNNs trained with different during-training calibration methods. The methods we use are mixup, focal loss, and label smoothing, in addition to the original training with cross-entropy loss. We can see in Figure 4 that our proposed during-training calibration

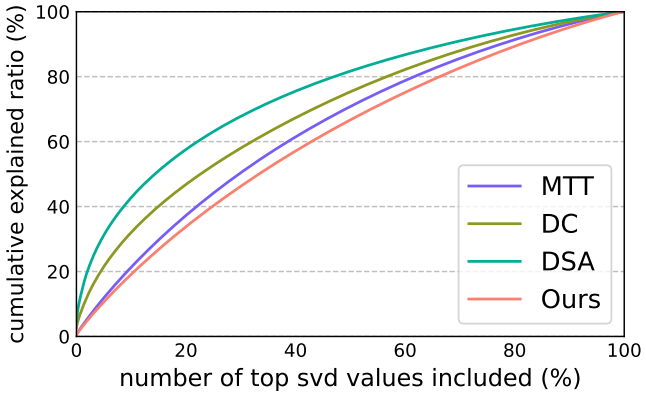


Figure 1. Cumulative *explained ratio*, i.e., percentage of top singular values to $\sum \text{diag}(\Sigma)$ in SVD decomposition of distilled CIFAR10 from different distillation backbones. Ours (red) grows at the most steady rate, indicating its evenly distributed information, compared to others with condensed information.

Figure 2. Ours (MTS) better calibrates DDNNs across different IPC in MTT. Left: CIFAR10. Right: CIFAR100.

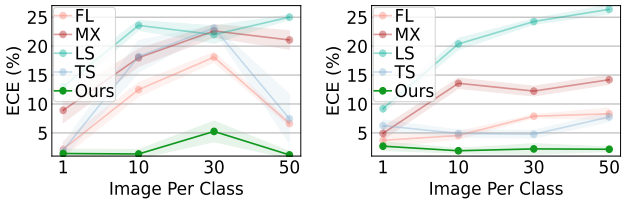
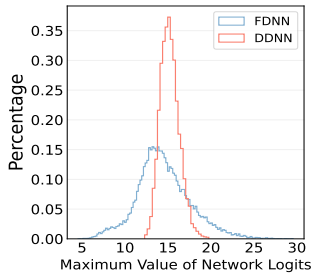


Figure 3. The more calibratable FDNN outputs more evenly distributed logits, while the less calibratable DDNN outputs a more concentrated logit distribution.



(a)

Table 2. ECE (%) of different calibration methods on FDNNs. With a low masking ratio r, our results (shadow) are comparable to temperature scaling and most of the time beats other methods. As our method is specifically designed for DDNNs, in the case of FDNNs where traditional methods are suitable, we can simply convert our method to temperature scaling by setting r to 0.

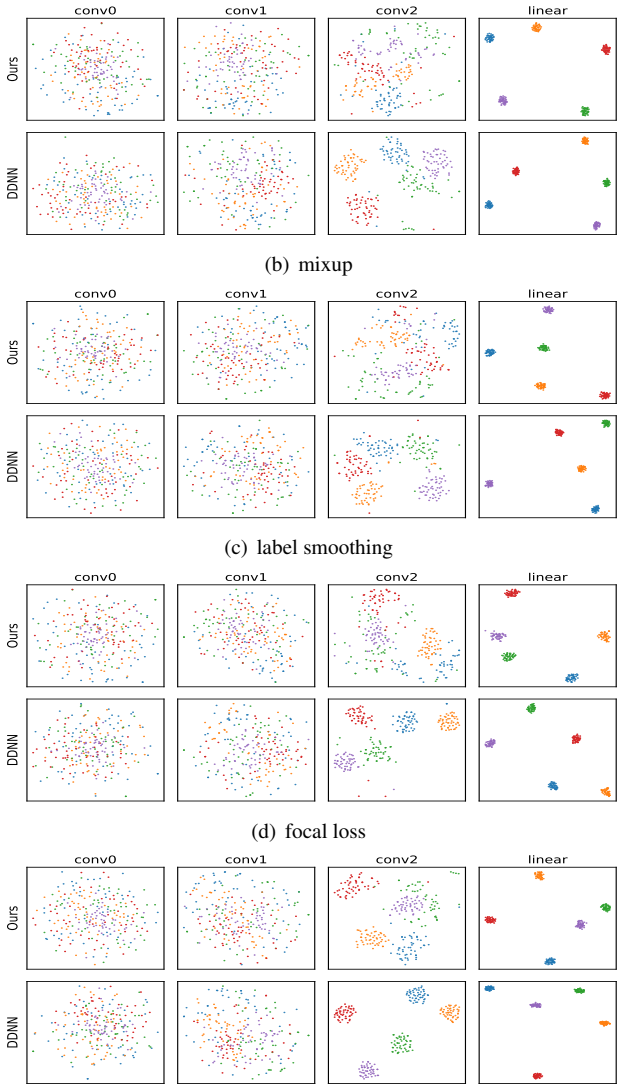
Dataset	Raw	TS	MX	LS	FL	MTS
CIFAR10	4.50	0.99	14.80	11.85	1.78	2.67
CIFAR100	13.05	1.41	10.69	7.17	3.49	1.84
Tiny ImageNet	22.26	4.95	6.34	3.29	12.55	4.93
ImageNette	10.90	2.81	11.22	22.24	5.21	2.87

MDT alleviates the issue of concentrate features for all the traditional methods used, giving better encoding potentials of DDNNs for transfer learning tasks, which leads to more calibratable DDNNs.

B.4. More Results on CIFAR100: ECE on different IPCs, max logits

We show in Figure2 that our MTS outperforms others in ECE on different IPCs. In the main paper, we mainly present IPC = 10 on Tiny-ImageNet & Subsets with MTT, 10 on CIFAR100 with DC/DSA (released), and 50 on others. These DD settings have higher accuracy and would better represent real-world settings.

We also provide visualization of maximum logits of DDNN on original MTT in Figure 3, in addition to the results on CIFAR10 in our main paper.



(b) mixup

(c) label smoothing

(d) focal loss

(e) no calibration (original cross-entropy)

Figure 4. T-SNE projections of feature vectors from each layer of a 4-block ConvNet trained with mixup, label smoothing, focal loss, and the original cross-entropy on distilled CIFAR10. In each training method, applying our proposed MDT (Ours) helps the network encode more source information in intermediate layers, as visualized by the rich features not separated until the last layer. The original DDNN poorly encodes source information, as shown by the feature projections already separated in layer conv2.

B.5. Performance Analysis of FDNNs

We further test MTS on the more calibratable FDNNs. We calibrate networks trained on the full CIFAR10, CIFAR100, TinyImageNet, and Nette subset of ImageNet. We report the mean of 2 runs due to limited computational resources. As reported in Table 2, our method performs comparably with existing well-developed methods. In realistic settings with a large amount of training data, we can set the masking ratio r to 0, which converts the MTS back to normal temperature scaling.

References

[1] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4750–4759, 2022. 1

[2] Zhiwei Deng and Olga Russakovsky. Remember the past: Distilling datasets into addressable memories for neural networks. In *Neural Information Processing Systems (NeurIPS)*, 2022. 1

[3] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017. 1

[4] Jeremy Howard. Imagenette: A smaller subset of 10 easily classified classes from imagenet, March 2019. 1

[5] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. 1

[6] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 1

[7] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010. 1

[8] Edgar Riba, Dmytro Mishkin, Daniel Ponsa, Ethan Rublee, and Gary Bradski. Kornia: an open source differentiable computer vision library for pytorch. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3674–3683, 2020. 1

[9] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 1

[10] Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *International Conference on Machine Learning*, pages 12674–12685. PMLR, 2021. 1

[11] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. *arXiv preprint arXiv:2006.05929*, 2020. 1