Reinforce Data, Multiply Impact: Improved Model Accuracy and Robustness with Dataset Reinforcement

Fartash Faghri*, Hadi Pouransari, Sachin Mehta, Mehrdad Farajtabar,
Ali Farhadi, Mohammad Rastegari, Oncel Tuzel
Apple

Abstract

We propose Dataset Reinforcement, *a strategy to improve a dataset once such that the accuracy of any model architecture trained on the reinforced dataset is improved at no additional training cost for users. We propose a Dataset Reinforcement strategy based on data augmentation and knowledge distillation. Our generic strategy is designed based on extensive analysis across CNN- and transformer-based models and performing large-scale study of distillation with state-of-the-art models with various data augmentations. We create a reinforced version of the ImageNet training dataset, called ImageNet$^+$, as well as reinforced datasets CIFAR-100$^+$, Flowers-102$^+$, and Food-101$^+$. Models trained with ImageNet$^+$ are more accurate, robust, and calibrated, and transfer well to downstream tasks (e.g., segmentation and detection). As an example, the accuracy of ResNet-50 improves by 1.7% on the ImageNet validation set, 3.5% on ImageNetV2, and 10.0% on ImageNet-R. Expected Calibration Error (ECE) on the ImageNet validation set is also reduced by 9.9%. Using this backbone with Mask-RCNN for object detection on MS-COCO, the mean average precision improves by 0.8%. We reach similar gains for MobileNets, ViTs, and Swin-Transformers. For MobileNetV3 and Swin-Tiny, we observe significant improvements on ImageNet-R/A/C of* ***up to 20% improved robustness****. Models pretrained on ImageNet$^+$ and fine-tuned on CIFAR-100$^+$, Flowers-102$^+$, and Food-101$^+$, reach up to 3.4% improved accuracy. The code, datasets, and pretrained models are available at https://github.com/apple/ml-dr.*

1. Introduction

With the advent of the CLIP [47], the machine learning community got increasingly interested in massive datasets whereby the models are trained on hundreds of millions of samples, which is orders of magnitude larger than the conventional ImageNet [15] with 1.2M samples. At the same time, models have gradually grown larger in multiple domains [1]. In computer vision, the state-of-the-art models have upwards of 300M parameters according to the Timm [63] library (e.g., BEiT [3], DeiT III [60], ConvNeXt [39]) and process inputs

*Correspondence to fartash@apple.com.

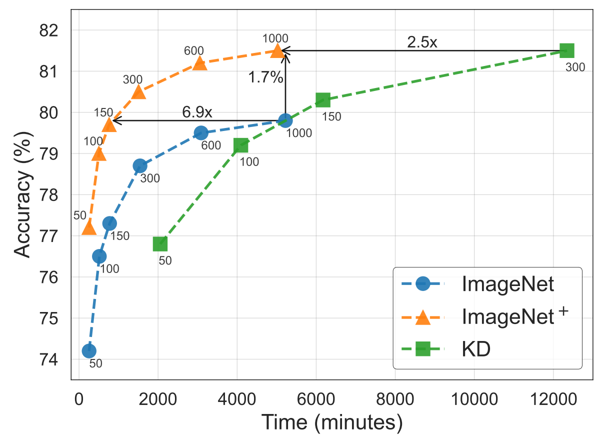


Figure 1: **Reinforced ImageNet, ImageNet$^+$, improves accuracy at similar iterations/wall-clock.** ImageNet validation accuracy of ResNet-50 is shown as a function of training duration with (1) ImageNet dataset, (2) knowledge distillation (KD), and (3) ImageNet$^+$ dataset (ours). Each point is a full training with epochs varying from 50-1000. An epoch has the same number of iterations for ImageNet/ImageNet$^+$.

Model	+Data Augmentation	+Reinforced Dataset(s)	ImageNet	CIFAR-100	Flowers-102	Food-101
MobileNetV3-Large	✗	✗	75.8	84.4	92.5	86.1
	✗	✓	**77.9**	**87.5**	**95.3**	**89.5**
ResNet-50	RandAugment	✗	80.4	88.4	93.6	90.0
	AutoAugment	✗	80.2	87.9	95.1	89.0
	TrivialAugWide	✗	80.4	87.9	94.8	89.3
	✗	✓	**82.0**	**89.8**	**96.3**	**92.1**
SwinTransformer-Tiny	RandAugment	✗	81.3	90.7	96.3	92.3
	✗	✓	**84.0**	**91.2**	**97.0**	**92.9**

Table 1: **Training/fine-tuning on reinforced datasets improve accuracy for a variety of architectures.** We reinforce each dataset *once* and train multiple models with similar cost as training on the original dataset. For datasets other than ImageNet, we fine-tune ImageNet/ImageNet$^+$ pretrained models. Dataset reinforcement significantly benefits from efficiently reusing the knowledge of a teacher.

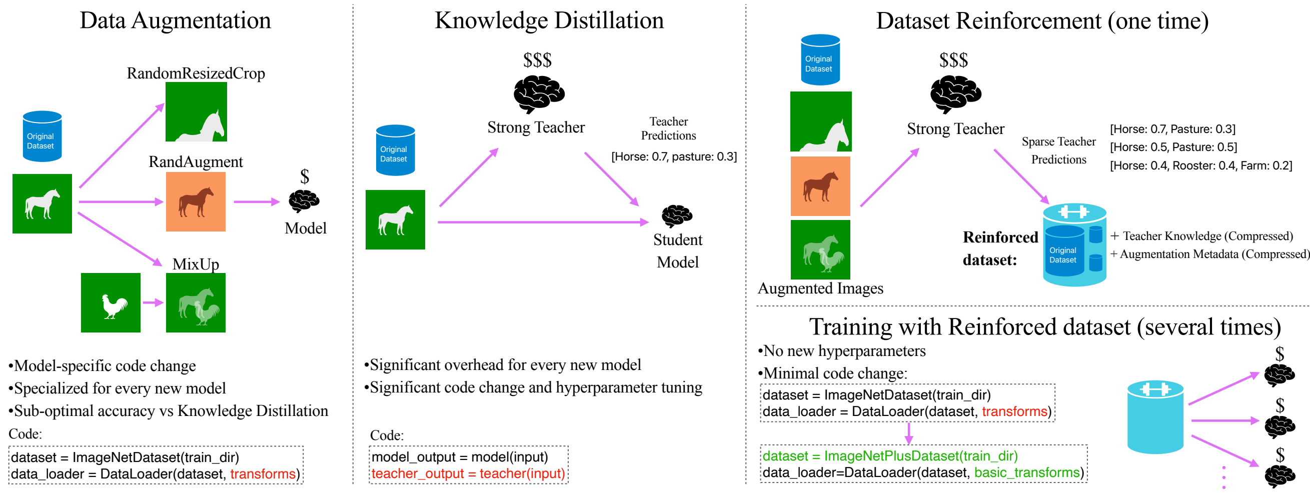


Figure 2: **Illustration of Dataset Reinforcement.** Data augmentation and knowledge distillation are common approaches to improving accuracy. Dataset reinforcement combines the benefits of both by bringing the advantages of large models trained on large datasets to other datasets and models. Training of new models with a reinforced dataset is as fast as training on the original dataset for the same total iterations. Creating a reinforced dataset is a one-time process (e.g., ImageNet to ImageNet^{+}) the cost of which is amortized over repeated uses.

at up to 800×800 resolution (e.g., EfficientNet-L2-NS [65]). Recent multi-modal vision-language models have up to 1.9B parameters (e.g., BeiT-3 [62]).

On the other side, there is a significant demand for small models that satisfy stringent hardware requirements. Additionally, there are plenty of tasks with small datasets that are challenging to scale because of the high cost associated with collecting and annotating new data. We seek to bridge this gap and bring the benefits of large models to any large, medium, or small dataset. We use knowledge from large models [47, 16, 7] to enhance the training of new models.

In this paper, we introduce *Dataset Reinforcement (DR)* as a strategy that improves the accuracy of models through reinforcing the training dataset. Compared to the original training data, a method for dataset reinforcement should satisfy the following desiderata:

- **No overhead for users**: Minimal increase in the computational cost of training a new model for similar total iterations (e.g., similar wall-clock time and CPU/GPU utilization).
- **Minimal changes in user code and model**: Zero or minimal modification to the training code and model architecture for the users of the reinforced dataset (e.g., only the dataset path and the data loader need to change).
- **Architecture independence**: Improve the test accuracy across variety of model architectures.

To understand the importance of the DR desiderata, let us discuss two common methods for performance improvements: data augmentation and knowledge distillation. Illustration in Fig. 2 compares these methods and our strategy for dataset reinforcement.

Data augmentation is crucial to the improved performance of machine learning models. Many state-of-the-art vision models [21, 27, 25] use the standard Inception-style augmentation [57] (i.e., random resized crop and random horizontal flipping) for training. In addition to these standard augmentation methods, recent models [59, 38] also incorporate mixing augmentations (e.g., MixUp [72] and CutMix [70]) and automatic augmentation methods (e.g., RandAugment [14] and AutoAugment [13]) to generate new data. However, data augmentation fails to satisfy all the desiderata as it does not provide architecture independent generalization. For example, light-weight CNNs perform best with standard Inception-style augmentations [25] while vision transformers [59, 38] prefer a combination of standard as well as advanced augmentation methods.

Knowledge distillation (KD) refers to the training of a student model by matching the output of a teacher model [35]. KD has consistently been shown to improve the accuracy of new models independent of their architecture significantly more than data augmentations [59]. However, knowledge distillation is expensive as it requires performing the inference (forward-pass) of an often significantly large teacher model at every training iteration. KD also requires modifying the training code to perform two forward passes on both the teacher and the student. As such, KD fails to satisfy minimal overhead and code change desiderata.

This paper proposes a dataset reinforcement strategy that exploits the advantages of both knowledge distillation and data augmentation by removing the training overhead of KD and finding generalizable data augmentations. Specifically, we introduce the *ImageNet*$^+$ dataset that provides a balanced trade-off between accuracies on a variety of models and has the same wall-clock as training on ImageNet for the same number of iterations (Fig. 1 and Tab. 1). To train models using the ImageNet$^+$ dataset, one only needs to change a few lines of the user code to use a modified data loader that reinforces every sample loaded from the training set.

Summary of contributions:

- We present a comprehensive large scale study of knowledge distillation from 80 pretrained state-of-the-art models and their ensembles. We observe that ensembles of state-of-the-art models trained on massive datasets generalize across student architectures (Sec. 2.1).
- We reinforce ImageNet by efficiently storing the knowledge of a strong teacher on a variety of augmentations. We investigate the generalizability of various augmentations for dataset reinforcement and find a tradeoff controlled by the reinforcement difficulty and model complexity (Sec. 2.2). This tradeoff can further be alleviated using curriculums based on the reinforcements (Appendix C.4).
- We introduce ImageNet$^+$, a reinforced version of ImageNet, that provides up to 4% improvement in accuracy for a variety of architectures in short as well as long training. We show that ImageNet$^+$ pretrained models result in 0.6-0.8 improvements in mAP for detection on MS-COCO and 0.3-1.3% improvement in mIoU for segmentation on ADE-20K (Sec. 3.1).
- Similarly, we create CIFAR-100$^+$, Flowers-102$^+$, and Food-101$^+$, and demonstrate their effectiveness for fine-tuning (Sec. 2.3). ImageNet$^+$ pretrained models fine-tuned on CIFAR-100$^+$, Flowers-102$^+$, and Food-101$^+$ show up to 3% improvement in transfer learning on CIFAR-100, Flowers-102, and Food-101.
- To further investigate this emergent transferablity we study robustness and calibration of the ImageNet$^+$ trained models. They reach up to 20% improvement on a variety of OOD datasets, ImageNet-(V2, A, R, C, Sketch), and ObjectNet (Sec. 3.2). We also show that models trained on ImageNet$^+$ are well calibrated compared to their non-reinforced alternatives (Sec. 3.3).

Our ImageNet$^+$, CIFAR-100$^+$, Flowers-102$^+$, and Food-101$^+$ reinforcements along with code to reinforce new datasets are available at https://github.com/apple/ml-dr.

2. Dataset Reinforcement

Our proposed strategy for dataset reinforcement (DR) is efficiently combining knowledge distillation and data augmentation to generate an enhanced dataset. We precompute and store the output of a strong pretrained model on multiple augmentations per sample as reinforcements. The stored outputs are more informative and useful for training compared with ground truth labels. This approach is related to prior works, such as Fast Knowledge Distillation (FKD) [55]) and ReLabel [71], that aim to improve the labels. Beyond these works, our goal is to find generalizable reinforcements that improve the accuracy of any architecture. First we perform a comprehensive study to find a strong teacher (Sec. 2.1) then find generalizable reinforcements on ImageNet (Sec. 2.2). To demonstrate the generality of our strategy and findings, we further reinforce CIFAR-100, Flowers-102, and Food-101 (Sec. 2.3).

The reinforced dataset consists of the original dataset plus the reinforcement meta data for all training samples. During the reinforcement process, for each sample a fixed number of reinforcements is generated using parametrized augmentation operations and evaluating the teacher predictions. To save storage, instead of storing the augmented images, the augmentation parameters are stored alongside the sparsified output of the teacher. As a result, the extra storage needed is only a fraction of the original training set for large datasets. Using our reinforced dataset has no computational overhead on training, requires no code change, and provides improvements for various architectures.

2.1. What is a good teacher?

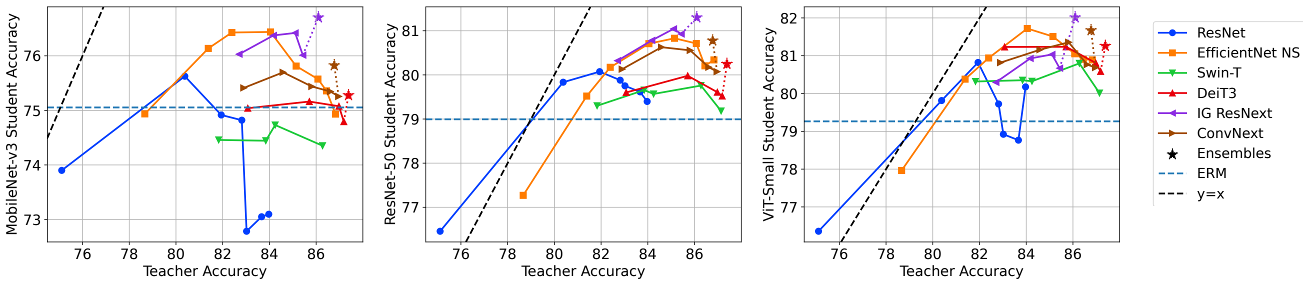
Knowledge distillation (KD) refers to training a student model using the outputs of a teacher model [9, 2, 35]. The training objective is as follows:

$$\min_{\theta} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D},\hat{\boldsymbol{x}}\sim\mathcal{A}(\boldsymbol{x})}\mathcal{L}(f_\theta(\hat{\boldsymbol{x}}), g(\hat{\boldsymbol{x}}))\,, \tag{1}$$

where, $\mathcal{D}$ is the training dataset, $\mathcal{A}$ is augmentation function, f_θ is the student model parameterized with θ, g is the teacher model, and $\mathcal{L}$ is the loss function between student and teacher outputs. Throughout this paper, we use the KL loss without a temperature hyperparameter and no mixing with the cross-entropy loss. We teach the student to imitate the output of the teacher on all augmentations consistent with [6].

It is common to use a fixed teacher because repeating experiments and selecting the best teacher is expensive [6, 19]. The teacher is often selected based on the state-of-the-art test accuracy of available pretrained models. However, it has been observed that most accurate models do not necessarily appear to be the best teachers [12, 43]. Ensemble models on the other hand, have been shown to be promising teachers from the early work of [9] until recent works in various domains [10, 68, 54, 56] and with techniques to boost the their performance [52, 17, 41]. None of these works have comprehensively studied finding the best teacher along with the necessary augmentations that result in consistent improvements over multiple student architectures.

To understand what makes a good teacher to reinforce datasets, we perform knowledge distillation with a variety of



(a) Light-weight CNN (MobileNetV3) (b) Heavy-weight CNN (ResNet-50) (c) Transformer (ViT-Small)

Figure 3: **Knowledge Distillation with models and ensembles from Timm library.** We observe the validation accuracy of students saturates or drops as the accuracy of teachers within an architecture family increases. We also observe that ensembles (marked by asterisks) are better teachers. Ensemble of IG-ResNext models performs best as teachers across student architectures. ERM (Empirical Risk Minimization) is standard training without knowledge distillation. Similar results for 150 epoch training in Fig. 7.

pretrained models in the Timm library [63] distilled to three representative student architectures MobileNetV3-large [25], ResNet-50 [21], and ViT-Small [16]. MobileNetV3 represents light-weight CNNs that often prefer easier training. ResNet-50 represents heavy-weight CNNs that can benefit from difficult training regimes but do not heavily rely on it because of their implicit inductive bias of the architecture. ViT-small represents the transformer architectures that have less implicit bias compared with CNNs and learn better in the presence of complex and difficult datasets. We consider various families of models as teachers including ResNets (34–152 and type d variants) [21], ConvNeXt family pretrained on the ImageNet-22K and fine-tuned on ImageNet-1K [39], DeiT-3 pretrained on the ImageNet-21K and fine-tuned on ImageNet-1K, IG-ResNext pretrained on the Instagram dataset [40], EfficientNets with Noisy Student training [65], and Swin-TransformersV2 pretrained with and without ImageNet-22K and fine-tuned on ImageNet-1K [37]. This collection covers a variety of vision transformers and CNNs pretrained on a wide spectrum of dataset sizes. We train all students with 224×224 inputs and follow [6] to match the resolution of teachers optimized to take larger inputs by passing the large crop to the teacher and resize it to 224×224 for the student.

We present the accuracies of students trained for 300 epochs as a function of the teacher accuracy in Fig. 3. Focusing first on the single (non-ensemble) networks (marked by circles), consistent with prior work, we observe that the most accurate models are not usually the best teachers [43]. For CNN model families (ResNets, EfficientNets, ResNexts, and ConvNeXts), the student accuracy is generally correlated with the teacher accuracy. When increasing the teacher accuracy, the student first improves but then it starts to saturate or even drops with the most accurate member of the family. Vision Transformers (Swin-Transformers, and DeiT-3) as teachers do not show the same trend as the accuracy of the students flattens across different teachers. Recently,[36] suggested that temperature tuning can help in KD from larger teachers. We do not adopt such hyperparameter tuning strategies in favor of architecture-independence and generalizability of dataset reinforcement.

On the other side, ensembles of state-of-the-art models (marked by asterisks) are consistently better teachers compared with any individual member of the family. We create 4-member ensembles of the best models from IG-ResNexts, ConvNeXts, and DeiT3 to cover CNNs, vision transformers, and extra data models. We find IG-ResNext teacher to provide a balanced improvement across all students. IG-ResNext models are also trained with 224×224 inputs while, for example, the best teacher from EfficientNet-NS family, EfficientNet-L2-NS, performs best at larger resolutions that is significantly more expensive to train with.

One of the benefits of dataset reinforcement paradigm is that the teacher can be expensive to train and use as long as we can afford to run it *once* on the target dataset for reinforcement. Also, the process of dataset reinforcement is highly parallelizable because performing the forward-pass on the teacher to generate predictions on multiple augmentations does not depend on any state or any optimization trajectory. For these reasons, we also considered significantly scaling knowledge distillation to super large ensembles with up to 128 members. We discuss our findings in Appendix B.2. Full table of accuracies for this section are in Appendix B.1.

2.2. ImageNet$^+$: What is the best combination of reinforcements?

In this section, we introduce ImageNet$^+$, a reinforcement of ImageNet. We create ImageNet$^+$ using the IG-ResNext ensemble (Sec. 2.1). Following [55], we store top 10 sparse probabilities for 400 augmentations per training sample in the ImageNet dataset [15]. We consider the following augmentations: Random-Resize-Crop (*RRC*), MixUp [72] and CutMix [70] (*Mixing*), and RandomAugment [14] and RandomErase (*RA/RE*). We also combine *Mixing* with *RA/RE*

	Sparse teacher prob.	Random Resize Crop + Horizontal Flip	Random Augment + Random Erase	MixUp + CutMix
ImageNet$^+$ variant	All	All	+RA/RE, +M*+R*	+Mixing, M*+R*
Apply probability	1	1, 0.5	1, 0.25	0.5, 0.5
Parameters	10× (Index, Prob)	4× Coords + Flip bit	2× (Op Id, Magnitude) + 4× Coords	(Img Id, λ) + (Img Id, 4× Coords)
Storage space (in bytes)	$10 \times (2 \times 4)$	$4 \times 4 + 1$	$2 \times 2 \times 4 + 4 \times 4$	$2 \times 4 + (1 + 4) \times 4$
Total storage space (400 samples per image)	38 GB	8 GB	15 GB	13 GB

Table 2: **Additional storage in ImageNet$^+$ variants**. Total additional storage for ImageNet$^+$ (*RRC+RA/RE*) is 61 GBs.

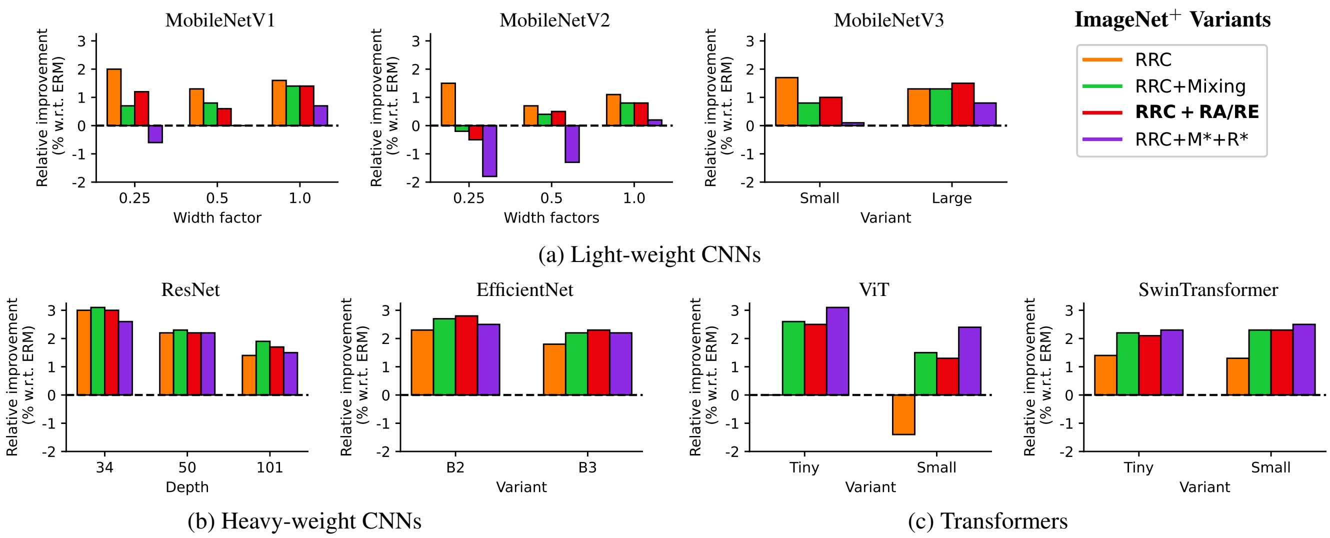


Figure 4: **Improvements across architectures with ImageNet$^+$ variants compared with ImageNet.** Top-1 accuracy of different models on the ImageNet validation set consistently improves when trained with the proposed datasets as compared to the standard ImageNet training set (Epochs=150). Our proposed dataset variant with *RRC+RA/RE*, **ImageNet$^+$**, provides balanced improvements of 1-4% across architectures. Further improvements with longer training (300-1000 epochs) in Tab. 4.

and refer to it as M^*+R^*. We add all augmentations on top of *RRC* and for clarity add + as shorthand for *RRC*+. We provide a summary of the reinforcement data stored for each ImageNet$^+$ variant in Tab. 2.

Models We study light-weight CNN-based (MobileNetV1 [26]/ V2 [50]/ V3[25]), heavy-weight CNN-based (ResNet [21] and EfficientNet [58]), and transformer-based (ViT [16] and SwinTransformer [38]) models. We follow [42, 64] and use state-of-the-art recipes, including optimizers, hyperparameters, and learning schedules, specific to each model on the ImageNet. We perform **no hyperparameter tuning specific to ImageNet$^+$** and achieve improvements with the same setup as ImageNet for all models.

Better accuracy We evaluate the performance of each model in terms of top-1 accuracy on the ImageNet validation set. Figure 4 compares the performance of different models trained using ImageNet and ImageNet$^+$ datasets. Fig. 4a shows that light-weight CNN models do not benefit from difficult reinforcements. This is expected because of their limited capacity. On the other side, both heavy-weight CNN (Fig. 4b) and transformer-based (Fig. 4c) models benefit from difficult reinforcements (*RRC+Mixing*, *RRC+RA/RE*, and *RRC+M^* +R^**). However, transformer-based models deliver best performance with the most difficult reinforcement (*RRC+M^* +R^**). This concurs with previous works that show transformer-based models, unlike CNNs, benefit from more data regularization as they do not have inductive biases [16, 59].

Overall, ***RRC+RA/RE*** provides a balanced trade-off between performance and model size across different models. Therefore, in the rest of this paper, we use *RRC+RA/RE* as our reinforced dataset and call it **ImageNet$^+$**. In the rest of the paper, we show results for three models that spans different model sizes and architecture designs (MobileNetV3-Large, ResNet-50, and SwinTransformer-Tiny).

We note that our observations are consistent across different architectures and recommend to see Appendix A for comprehensive results on 25 architectures. We provide expanded ablation studies in Appendix C using a cheaper teacher, ConvNext-Base-IN22FT1K. For example, we find 1) The number of stored samples can be 3× fewer than intended training epochs, 2) Additional augmentations on top of ImageNet$^+$ are not useful. 3) Tradeoff in reinforcement difficulty can be further reduced with curriculums. 4) Curriculums are better than various sample selection methods at the time of reinforcing the dataset. We provide all hyperparameters and training recipes in Appendix G.

Pretraining Dataset	CIFAR-100		Flowers-102		Food-101	
	Orig.	+	Orig.	+	Orig.	+
None	80.2	83.6	68.8	87.5	85.1	88.2
ImageNet	84.4	87.2	92.5	94.1	86.1	89.2
ImageNet$^+$ (Ours)	86.0	**87.5**	93.7	**95.3**	86.6	**89.5**

Table 3: **Pretraining and fine-tuning on reinforced datasets is up to 3.4% better than using non-reinforced datasets.** Top-1 accuracy on the test set for MobileNetV3-Large is shown. On Food-101, 86.1% is improved to 89.5%, demonstrating composition of reinforced datasets.

2.3. CIFAR-100$^+$, Flowers-102$^+$, Food-101$^+$: How to reinforce other datasets?

We reinforced ImageNet due to its popularity and effectiveness as a pretraining dataset for other tasks (e.g., object detection). Our findings on ImageNet are also useful for reinforcing other datasets and reduce the need for exhaustive studies. Specifically, we suggest the following guidelines: 1) use ensemble of strong teachers trained on large diverse data 2) balance reinforcement difficulty and model complexity.

In this section, we extend dataset reinforcement to three other datasets, CIFAR-100 [31], Flowers-102 [45], and Food-101 [8], with 50K, 1K, and 75K training data respectively. We build a teacher for each dataset by fine-tuning ImageNet$^+$ pretrained ResNet-152 that reaches the accuracy of 90.6%, 96.6%, and 91.8%, respectively. By repeating fine-tuning 4 times, we get three teacher ensembles of 4xResNet-152. Next we generate reinforcements using similar augmentations to ImageNet$^+$, that is *RRC+RA/RE*. We store 800, 8000, and 800 augmentations per original sample. After that, we train various models on the reinforced data at similar training time to standard training. To achieve the best performance, we use pretrained models on ImageNet/ImageNet$^+$ and fine-tune on each dataset for varying epochs up to 1000, 10000, and 1000 (for CIFAR-100, Flowers-102, and Food-101, respectively) and report the best result.

Table 3 shows that MobileNetV3-Large pretrained and fine-tuned with reinforced datasets reaches up to 3% better accuracy. We observe that pretraining and fine-tuning on reinforced datasets together give the largest improvements. We provide results for other models in Appendix D.

3. Experiments

Baseline methods We compare the performance of models trained using ImageNet$^+$ with the following baseline methods: (1) *KD* [35, 6] (Online distillation): A standard knowledge distillation method with strong teacher models and model-specific augmentations, (2) *MEALV2* [54] (Fine-tuning distillation): Distill knowledge to student with good initialization from multiple teachers, (3) *FunMatch* [6] (Patient online distillation): Distill for significantly many epochs with strong augmentations, (4) *ReLabel* [71] (Offline label-map distillation): Pre-computes global label maps from the pre-trained teacher, and (5) *FKD* [55] (Offline distillation): Pre-computes soft labels using multi-crop knowledge distillation. We consider FKD as the baseline approach for dataset reinforcement.

Model	Dataset	Training Epochs		
		150	300	1000
MobileNetV3-Large	ImageNet	74.7	74.9	75.1
	ImageNet$^+$ (Ours)	**76.2**	**77.0**	**77.9**
ResNet-50	ImageNet	77.4	78.8	79.6
	ImageNet$^+$ (Ours)	**79.6**	**80.6**	**81.7**
SwinTransformer-Tiny	ImageNet	79.9	80.9	80.9
	ImageNet$^+$ (Ours)	**82.0**	**83.0**	**83.8**

Table 4: **ImageNet$^+$ models consistently outperform ImageNet models when trained for longer**. Top-1 accuracy on the ImageNet validation set is shown. An epoch has the same number of iterations for ImageNet/ImageNet$^+$.

Longer training Recent works have shown that models trained for few epochs (e.g., 100 epochs) are sub-optimal and their performance improves with longer training [64, 16, 59]. Following these works, we train different models at three epoch budgets, i.e., 150, 300, and 1000 epochs, using both ImageNet and ImageNet$^+$ datasets. Table 4 shows models trained with ImageNet$^+$ dataset consistently deliver better accuracy in comparison to the ones trained on ImageNet. An epoch of ImageNet$^+$ consists of exactly one random reinforcement per sample in ImageNet.

Training and reinforcement time Table 4 shows ImageNet$^+$ improves the performance of various models. A natural question that arises is: *Does ImageNet$^+$ introduce computational overhead when training models?* On average, training MobileNetV3-Large, ResNet-50, and SwinTransformer-Tiny is $1.12\times$, $1.01\times$, and $0.99\times$ the total training time on ImageNet. The extra time for MobileNetV3 is because there is no data augmentations in our baseline. ImageNet$^+$ took 2205 GPUh to generate using 64xA100 GPUs, which is highly parallelizable. For comparison, training ResNet-50 for 300 epochs on 8xA100 GPUs takes 206 GPUh. The reinforcement generation is a one-time cost that is amortized over many uses. The time to reinforce other datasets and the storage is discussed in Appendix F.

Comparison with state-of-the-art methods Table 5 compares the performance of models trained with ImageNet$^+$ and existing methods. We make following observations: (1) Compared to the closely related method, i.e., FKD, models trained using ImageNet$^+$ deliver better accuracy. (2) We achieve comparable results to online distillation methods (e.g., FunMatch), but with fewer epochs and faster training (Fig. 1). (3) Small variants of the same family trained with ImageNet$^+$ achieve similar performance to larger models trained with ImageNet dataset. For example, ResNet-50

(81.7%) with ImageNet$^+$ achieves similar performance as ResNet-101 with ImageNet (81.5%). We observe similar phenomenon across other models, including light-weight CNN models. This enables replacing large models with smaller variants in their family for faster inference across devices, including edge devices, without sacrificing accuracy.

3.1. Transfer Learning

To evaluate the transferability of models pre-trained using ImageNet$^+$ dataset, we evaluate on following tasks: (1) semantic segmentation with DeepLabv3 [11] on the ADE20K dataset [74], (2) object detection with Mask-RCNN [20] on the MS-COCO dataset [34], and (3) fine-grained classification on the CIFAR-100 [31], Flowers-102 [45], and Food-101 [8] datasets.

Tables 6 and 8 show models trained on the ImageNet$^+$ dataset have better transferability properties as compared to the ImageNet dataset across different tasks (detection, segmentation, and fine-grained classification). To analyze the isolated impact of ImageNet$^+$ in this section, the fine-tuning datasets are not reinforced. We present all combinations of training with reinforced/non-reinforced pretraining/fine-tuning datasets in Appendix D.

Model	Dataset	Offline KD?	Random Init.?	Epochs	Accuracy
MobileNetV3 -Large	ImageNet [25]	NA	✓	600	75.2
	FunMatch [6]*	✗	✓	1200	76.3
	MEALV2 [54]	✗	✗	180	76.9
	ImageNet$^+$ (Ours)	✓	✓	300	**77.0**
ResNet-50	ImageNet [64]	NA	✓	600	80.4
	ReLabel [71]	✓	✓	300	78.9
	FKD [55]	✓	✓	300	80.1
	MEALV2 [54]	✗	✗	180	80.6
	ImageNet$^+$ (Ours)	✓	✓	300	80.6
	ImageNet$^+$ (Ours)	✓	✓	1000	**81.7**
	FunMatch [6]*	✗	✓	1200	**81.8**
ResNet-101	ImageNet [64]	NA	✓	1000	81.5
ViT-Tiny	ImageNet [59]	NA	✓	300	72.2
	DeiT [59]	✗	✓	300	74.5
	FKD [55]	✓	✓	300	75.2
	ImageNet$^+$ (Ours)	✓	✓	300	**75.8**
ViT-Small	ImageNet [59]	NA	✓	300	79.8
	DeiT [59]	✗	✓	300	81.2
	ImageNet$^+$ (Ours)	✓	✓	300	**81.4**
ViT-Base↑384	ImageNet [59]	NA	✓	300	83.1
	DeiT [59]	✗	✓	300	83.4
	ImageNet$^+$ (Ours)	✓	✓	300	**84.5**

Table 5: **Comparison with state-of-the-art methods on the ImageNet validation set.** Models trained with ImageNet$^+$ dataset deliver similar or better performance than existing methods. Importantly, unlike online KD methods (e.g., FunMatch or DeiT), ImageNet$^+$ does not add computational overhead to standard ImageNet training (Fig. 1). Here, NA denotes standard supervised ImageNet training with no online/offline KD. ↑384 denotes training at 384 resolution. An epoch has the same number of iterations for ImageNet/ImageNet$^+$.

Model	Pretraining dataset	Task	
		ObjDet	SemSeg
MobileNetV3-Large	ImageNet	35.5	37.2
	ImageNet$^+$ (Ours)	**36.1**	**38.5**
ResNet-50	ImageNet	42.2	42.8
	ImageNet$^+$ (Ours)	**42.5**	**44.2**
SwinTransformer-Tiny	ImageNet	45.8	41.2
	ImageNet$^+$ (Ours)	**46.5**	**42.5**

Table 6: **Transfer learning for object detection and semantic segmentation**. For object detection (ObjDet), we report standard mean average precision on MS-COCO dataset while for sementic segmentation (SemSeg), we report mean intersection accuracy on ADE20K dataset. Task datasets are not reinforced.

3.2. Robustness analysis

To evaluate the robustness of different models trained using the ImageNet$^+$ dataset, we evaluate on three subsets of the ImageNetV2 dataset [48], which is specifically designed to study the robustness of models trained on the ImageNet dataset. We also evaluate ImageNet models on other distribution shift datasets, ImageNet-A [24], ImageNet-R [22], ImageNet-Sketch [61], ObjectNet [4], and ImageNet-C [23]. We measure the top-1 accuracy except for ImageNet-C. On ImageNet-C, we measure the mean corruption error (mCE) and report 100 minus mCE.

Tab. 7 shows that models trained using ImageNet$^+$ dataset are up to 20% more robust. Overall, these robustness results in conjunction with results in Tab. 4 highlight the effectiveness of the proposed dataset.

3.3. Calibration: Why are ImageNet$^+$ models robust and transferable?

To understand why ImageNet$^+$ models are significantly more robust than ImageNet models we evaluate their Expected Calibration Error (ECE) [32] on the validation set. Fig. 5 shows that ImageNet$^+$ models are well-calibrated and significantly better than ImageNet models. This matches recent observations about ensembles that out-of-distribution robustness is better for well-calibrated models [33]. Full calibration results are presented in Appendix E.

3.4. Comparison with FKD and ReLabel.

We reproduce FKD and ReLabel with our training recipe as well as regenerate the dataset of FKD. We compare the accuracy on ImageNet validation and its distribution shifts as well as the cost of dataset generation/storage. We train models for 300 epochs.

Training recipe We report results of training with our code on the released datasets of ReLabel and FKD. In addition to reproducing FKD results by training on their released dataset of 500-sample per image, we also reproduce their dataset using our code and their teacher. Tab. 9 verifies that our

Model	Dataset	ImageNet-V2			ImageNet-A	ImageNet-R	ImageNet-Sketch	ObjectNet	ImageNet-C	Avg.
		V2-A	V2-B	V2-C						
MobileNetV3-Large	ImageNet	71.5	62.9	76.8	4.5	32.4	20.6	32.8	21.8	31.1
	ImageNet$^+$ (Ours)	**75.1**	**66.3**	**80.5**	**7.6**	**42.0**	**29.0**	**38.1**	**32.0**	**37.6**
ResNet-50	ImageNet	76.3	67.4	81.3	11.9	38.1	27.4	41.6	33.2	38.3
	ImageNet$^+$ (Ours)	**79.3**	**71.3**	**83.8**	**15.1**	**48.1**	**34.9**	**46.8**	**39.0**	**43.9**
SwinTransformer-Tiny	ImageNet	77.0	69.3	81.6	21.0	37.7	25.4	40.5	36.9	35.7
	ImageNet$^+$ (Ours)	**81.5**	**74.1**	**85.3**	**30.2**	**58.0***	**40.8**	**50.6**	**46.6**	**42.2**

Table 7: **ImageNet$^+$ models are up to 20% more robust on ImageNet distribution shifts**. All models are trained for 1000 epochs. We report on ImageNetV2 variations Threshold-0.7 (V2-A), Matched-Frequency (V2-B), and Top-Images (V2-C). We report accuracy on all datasets except for ImageNet-C where we report 100 minus mCE metric. * Largest improvement.

improvements are due to the superiority of ImageNet$^+$, not any other factors such as the training recipe. Our ImageNet$^+$-RRC is also closely related to FKD as it uses the same set of augmentations (random-resized-crop and horizontal flip) but together with our optimal teacher (4xIG-ResNext). We observe that ImageNet$^+$-RRC achieves better results than FKD but still lower than ImageNet$^+$ (Tab. 11c and Fig. 4).

Model	Pretraining dataset	Fine-tuning dataset		
		CIFAR-100	Flowers-102	Food-101
MobileNetV3-Large	ImageNet	84.4	92.5	86.1
	ImageNet$^+$ (Ours)	**86.0**	**93.7**	**86.6**
ResNet-50	ImageNet	88.4	93.6	90.0
	ImageNet$^+$ (Ours)	**88.8**	**95.0**	**90.5**
SwinTransformer-Tiny	ImageNet	90.6	96.3	92.3
	ImageNet$^+$ (Ours)	**90.9**	**96.6**	**93.0**

Table 8: **Transfer learning for fine-grained object classification.** Only pretraining dataset is reinforced and fine-tuning datasets are not reinforced. Reinforced pretraining/fine-tuning results in Tab. 1.

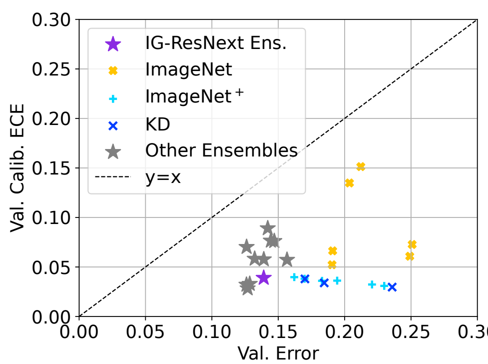


Figure 5: **ImageNet$^+$ models are well-calibrated**. We plot the Expected Calibration Error (ECE) on the ImageNet validation set over the validation error (normalized by 100 to range $[0, 1]$) for MobileNetV3/ResNet-50/Swin-Tiny architectures trained for 300 and 1000 epochs on ImageNet and ImageNet$^+$. ImageNet$^+$ models are significantly more calibrated, even matching or better than their teacher (IG-ResNext Ensemble). We also observe that the IG-ResNext model is one of the best calibrated models on the validation set from our pool of teachers.

Generation/Storage Cost We provide comparison of generation/storage costs in Tab. 9. In our reproduction, generating FKD's data takes 2260 GPUh, slightly more than ImageNet$^+$ because their teacher processes inputs at the larger resolution of 475×475 compared to our resolution of 224×224.

ImageNet$^+$-Small We subsampled ImageNet$^+$ into a variant that is 10.6 GBs, comparable to prior work. We reduce the number of samples per image to 100 and store teacher probabilities with top-5 sparsity. If not subsampled from ImageNet$^+$, generating ImageNet$^+$-Small would take half the time of FKD (200 samples) while still comparable in accuracy to ImageNet$^+$. Note that ImageNet$^+$ is more general-purpose and preferred, especially for long training.

3.5. CLIP-pretrained Teachers

In this section, we evaluate the performance of CLIP-pretrained models [47] fine-tuned on ImageNet as teachers. This study complements our large-scale study of teachers in Sec. 2.1 where we evaluated more than 100 SOTA large models and ensembles. Table 10 compares an ensemble of 4 CLIP-pretrained models to our selected ensemble of 4 IG-ResNext models as well as a mixture of ResNext, ConvNext, CLIP-ViT, and ViT (abrv. RCCV) models (See Appendix H for the model names). We generate new ImageNet$^+$ variants and train various architectures for 1000 epochs on each dataset. We observe that ImageNet$^+$ with our previously selected IG-ResNext ensemble is superior to CLIP-pretrained and mixed-architecture teachers across architectures. The CLIP variant provides near the maximum gain on Swin-Tiny and mixing it with IG-ResNext reduces the gap on CNNs.

4. Related work

We build on top of the well-known Knowledge Distillation framework [9, 2, 35], the effectiveness of which has been extensively studied [12, 56]. Numerous variants of KD have been proposed, including feature distillation [28, 73], iterative distillation [43, 67], and self-distillation [65, 44, 18, 29]. Label smoothing, an effective regularizer and related to KD, is particularly related to our work when interpreted as augmenting the output space [69, 53].

Dataset	Our Gen.	Our Train	Optimal Teacher	Optimal Aug.	Top-K	Num. Samples	Storage (GBs) Raw	Storage (GBs) GZIP	Gen. Time (GPUh)	ResNet-50 IN	ResNet-50 IN-OOD	Swin-Tiny IN	Swin-Tiny IN-OOD
ReLabel	✗	✓	✗	✗	5	1	10.7	4.8	10	79.5	45.7	81.2	48.2
FKD	✗	✓	✗	✗	5	200	13.6	8.9	904*	79.8	45.0	82.0	48.7
FKD	✗	✓	✗	✗	5	500	34.0	22.0	2260*	80.1	45.0	82.2	48.9
FKD	✓	✓	✗	✗	10	400	46.3	33.4	1808	79.8	45.0	82.1	49.0
ImageNet$^+$-RRC	✓	✓	✓	✗	10	400	46.3	33.4	1993	80.3	46.5	82.4	51.0
ImageNet$^+$-Small	✓	✓	✓	✓	5	100	10.6	5.6	551	**80.6**	**48.9**	**82.9**	**54.6**
ImageNet$^+$	✓	✓	✓	✓	10	400	61.5	37.5	2205	**80.6**	**49.1**	**83.0**	**54.7**

Table 9: **Comparison with Relabel and FKD. Up to 5.6% better than FKD on ImageNet-OOD**, the average of ImageNet-V2/A/R/S/O/C accuracies. Highlighted accuracies are within 0.2% of the best. Compared with prior work, we use an optimal teacher (4xIG-ResNext) and optimal combination of augmentations (RRC+RA/RE). * Our estimates.

Model	ImageNet	ImageNet$^+$ IG-ResNext*	ImageNet$^+$ CLIP	ImageNet$^+$ Mixed
MobileNetV3-Large	75.1	**77.9**$_{+2.9}$	77.2$_{+2.1}$	77.4$_{+2.3}$
ResNet-50	79.6	**81.7**$_{+2.1}$	81.1$_{+1.4}$	**81.5**$_{+1.8}$
Swin-Tiny	80.9	**83.8**$_{+2.8}$	**83.7**$_{+2.7}$	**83.8**$_{+2.8}$

Table 10: **Our selected IG-ResNext ensemble is superior to CLIP-pretrained ensembles.** We reinforce ImageNet dataset with an ensemble of CLIP-pretrained models as well as a mixture of multiple architectures and train various models for 1000 epochs. Subscripts show the improvement on top of the ImageNet accuracy. * Our chosen ImageNet$^+$ variant.

Closely related to our work, investigating and improving the accuracy on the ImageNet dataset has attracted much interest lately. [5] eliminated erroneous labeled examples in the training with reference to a strong classifier. In [51], ImageNet dataset evaluation was revisited and alternative test sets were released. Relabel [71] proposed storing multiple labels on various regions of an image using a teacher. FKD [55] further pushed this direction by caching the predictions of a strong teacher but with a limited augmentation. Similarly, in [49], the architecture-independent generalization of KD was exploited to propose a unified scheme for training with ImageNet seamlessly without any hyperparameter tuning or per-model training recipes. [36] identified the temperature hyperparameter in KD as an important factor limiting benefits of stronger augmentations and teachers, and proposed an adaptive scheme to dynamically set the temperature during training. Distilling feature maps and probability distributions between the random pair of original images and their MixUp images was proposed to guide the network to learn cross-image knowledge [46, 66]. For self-supervised learning, [30] adapted modern image-based regularizations with KD to improve the contrastive loss with some supervision. Our work has also been inspired by [6] where they proposed imitating the teacher on severe augmentations and train for thousands of epochs. With our proposed DR strategy, we significantly reduce the cost of function matching by storing a few samples and reusing them for longer training.

5. Conclusion

We go beyond the conventional online knowledge distillation and introduce Dataset Reinforcement (DR) as a general offline strategy. Our investigation unwraps tradeoffs in finding generalizable reinforcements controlled by the difficulty of augmentations and we propose ways to balance.

We study the choice of the teacher (more than 100 SOTA large models and ensembles), augmentation (4 more than prior work), and their impact on a diverse collection of models (25 architectures), especially for long training (up to 1000 epochs). We demonstrate significant improvements (up to 20%) in robustness, calibration and transfer (in/out of distribution classification, segmentation, and detection). Our novel method of training and fine-tuning on doubly reinforced datasets (e.g., ImageNet$^+$ to CIFAR-100$^+$) demonstrates new possibilities of DR as a generic strategy. We also study ideas that were not used in ImageNet$^+$, including curriculums, mixing augmentations and more in the appendix.

The proposed DR strategy is only an example of the large category of ideas possible within the scope of dataset reinforcement. Our desiderata would also be satisfied by methods that expand the training data, especially in limited data domains, using strong generative foundation models.

Limitations Limitations of the teacher can potentially transfer through dataset reinforcement. For example, overconfident biased teachers should not be used and diverse ensembles are preferred. Human verification of the reinforcements is also a solution. Note that original labels are unmodified in reinforced datasets and can be used in curriculums. Our robustness and transfer learning evaluations consistently show better transfer and generalization for ImageNet$^+$ models likely because of lower bias of the teacher ensemble trained on diverse data.

Acknowledgments

We would like to thank Arsalan Farooq, Farzad Abdolhosseini, Keivan Alizadeh-Vahid, Pavan Kumar Anasosalu Vasu, and Raviteja Vemulapalli for the enriching discussions. We also thank the reviewers for their valuable feedback.

References

[1] Ibrahim Alabdulmohsin, Behnam Neyshabur, and Xiaohua Zhai. Revisiting neural scaling laws in language and vision. *arXiv preprint arXiv:2209.06640*, 2022. 1

[2] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? *Advances in neural information processing systems*, 27, 2014. 3, 8

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022. 1

[4] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019. 7

[5] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020. 9

[6] Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10925–10934, 2022. 3, 4, 6, 7, 9, 19, 20

[7] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 2

[8] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 6, 7

[9] Cristian Buciluă, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006. 3, 8

[10] Yevgen Chebotar and Austin Waters. Distilling knowledge from ensembles of neural networks for speech recognition. In *Interspeech*, pages 3439–3443, 2016. 3

[11] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 7

[12] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4794–4802, 2019. 3, 8

[13] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018. 2

[14] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 2, 4, 19

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 4

[16] 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. *arXiv preprint arXiv:2010.11929*, 2020. 2, 4, 5, 6

[17] Rasool Fakoor, Jonas W Mueller, Nick Erickson, Pratik Chaudhari, and Alexander J Smola. Fast, accurate, and simple models for tabular data via augmented distillation. *Advances in Neural Information Processing Systems*, 33:8671–8681, 2020. 3

[18] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International Conference on Machine Learning*, pages 1607–1616. PMLR, 2018. 8

[19] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021. 3

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 7

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 4, 5

[22] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021. 7

[23] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019. 7

[24] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021. 7

[25] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 2, 4, 5, 7

[26] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 5

[27] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2

[28] Mingi Ji, Byeongho Heo, and Sungrae Park. Show, attend and distill: Knowledge distillation via attention-based fea-

ture matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7945–7952, 2021. 8

[29] Mingi Ji, Seungjae Shin, Seunghyun Hwang, Gibeom Park, and Il-Chul Moon. Refine myself by teaching myself: Feature refinement via self-knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10664–10673, 2021. 8

[30] Jaewon Kim, Jooyoung Chang, and Sang Min Park. A generalized supervised contrastive learning framework. *arXiv preprint arXiv:2206.00384*, 2022. 9

[31] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6, 7

[32] Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. *Advances in Neural Information Processing Systems*, 32, 2019. 7

[33] Ananya Kumar, Tengyu Ma, Percy Liang, and Aditi Raghunathan. Calibrated ensembles can mitigate accuracy tradeoffs under distribution shift. In *Uncertainty in Artificial Intelligence*, pages 1041–1051. PMLR, 2022. 7

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 7

[35] Yih-Kai Lin, Chu-Fu Wang, Ching-Yu Chang, and Hao-Lun Sun. An efficient framework for counting pedestrians crossing a line using low-cost devices: the benefits of distilling the knowledge in a neural network. *Multim. Tools Appl.*, 80(3):4037–4051, 2021. 2, 3, 6, 8

[36] Jihao Liu, Boxiao Liu, Hongsheng Li, and Yu Liu. Meta knowledge distillation. *arXiv preprint arXiv:2202.07940*, 2022. 4, 9

[37] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12009–12019, 2022. 4

[38] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2, 5

[39] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 4

[40] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018. 4

[41] Andrey Malinin, Bruno Mlodozeniec, and Mark Gales. Ensemble distribution distillation. *arXiv preprint arXiv:1905.00076*, 2019. 3

[42] Sachin Mehta, Farzad Abdolhosseini, and Mohammad Rastegari. Cvnets: High performance library for computer vision. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, 2022. 5, 13, 26

[43] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5191–5198, 2020. 3, 4, 8

[44] Hossein Mobahi, Mehrdad Farajtabar, and Peter Bartlett. Self-distillation amplifies regularization in hilbert space. *Advances in Neural Information Processing Systems*, 33:3351–3361, 2020. 8

[45] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 6, 7

[46] Hadi Pouransari, Mojan Javaheripi, Vinay Sharma, and Oncel Tuzel. Extracurricular learning: Knowledge transfer beyond empirical distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3032–3042, June 2021. 9

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. 1, 2, 8

[48] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 7

[49] Tal Ridnik, Hussam Lawen, Emanuel Ben-Baruch, and Asaf Noy. Solving imagenet: a unified scheme for training any backbone to top results. *arXiv preprint arXiv:2204.03475*, 2022. 9

[50] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 5

[51] Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. Evaluating machine accuracy on imagenet. In *International Conference on Machine Learning*, pages 8634–8644. PMLR, 2020. 9

[52] Zhiqiang Shen, Zhankui He, and Xiangyang Xue. Meal: Multi-model ensemble via adversarial learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4886–4893, 2019. 3

[53] Zhiqiang Shen, Zechun Liu, Dejia Xu, Zitian Chen, Kwang-Ting Cheng, and Marios Savvides. Is label smoothing truly incompatible with knowledge distillation: An empirical study. *arXiv preprint arXiv:2104.00676*, 2021. 8

[54] Zhiqiang Shen and Marios Savvides. Meal v2: Boosting vanilla resnet-50 to 80%+ top-1 accuracy on imagenet without tricks. *arXiv preprint arXiv:2009.08453*, 2020. 3, 6, 7

[55] Zhiqiang Shen and Eric Xing. A fast knowledge distillation framework for visual recognition. *arXiv preprint arXiv:2112.01528*, 2021. 3, 4, 6, 7, 9, 19

[56] Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A. Alemi, and Andrew Gordon Wilson. Does knowledge distillation really work? In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 6906–6919, 2021. 3, 8

[57] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 2

[58] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 5

[59] 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*, pages 10347–10357. PMLR, 2021. 2, 5, 6, 7

[60] Hugo Touvron, Matthieu Cord, and Herve Jegou. Deit iii: Revenge of the vit. *arXiv preprint arXiv:2204.07118*, 2022. 1

[61] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019. 7

[62] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 2

[63] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019. 1, 4, 16

[64] Ross Wightman, Hugo Touvron, and Hervé Jégou. Resnet strikes back: An improved training procedure in timm. *arXiv preprint arXiv:2110.00476*, 2021. 5, 6, 7, 26

[65] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020. 2, 4, 8

[66] Chuanguang Yang, Zhulin An, Helong Zhou, Linhang Cai, Xiang Zhi, Jiwen Wu, Yongjun Xu, and Qian Zhang. Mixskd: Self-knowledge distillation from mixup for image recognition. In *European Conference on Computer Vision*, pages 534–551. Springer, 2022. 9

[67] Chenglin Yang, Lingxi Xie, Chi Su, and Alan L Yuille. Snapshot distillation: Teacher-student optimization in one generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2859–2868, 2019. 8

[68] Shan You, Chang Xu, Chao Xu, and Dacheng Tao. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1285–1294, 2017. 3

[69] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3903–3911, 2020. 8

[70] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 2, 4, 19

[71] Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, Junsuk Choe, and Sanghyuk Chun. Re-labeling imagenet: From single to multi-labels, from global to localized labels. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 2340–2350. Computer Vision Foundation / IEEE, 2021. 3, 6, 7, 9

[72] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 2, 4, 19

[73] Linfeng Zhang and Kaisheng Ma. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *International Conference on Learning Representations*, 2020. 8

[74] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 7