Are Large Kernels Better Teachers than Transformers for ConvNets?

Tianjin Huang [1] **Lu Yin** [1] **Zhenyu Zhang** [2] **Li Shen** [3] **Meng Fang** [4 1]
Mykola Pechenizkiy [1] **Zhangyang Wang** [2] **Shiwei Liu** [2]

Abstract

This paper reveals a new appeal of the recently emerged large-kernel Convolutional Neural Networks (ConvNets): as the teacher in Knowledge Distillation (KD) for small-kernel ConvNets. While Transformers have led state-of-the-art (SOTA) performance in various fields with ever-larger models and labeled data, small-kernel ConvNets are considered more suitable for resource-limited applications due to the efficient convolution operation and compact weight sharing. KD is widely used to boost the performance of small-kernel ConvNets. However, previous research shows that it is not quite effective to distill knowledge (e.g., global information) from Transformers to small-kernel ConvNets, presumably due to their disparate architectures. We hereby carry out a first-of-its-kind study unveiling that modern large-kernel ConvNets, a compelling competitor to Vision Transformers, are remarkably more effective teachers for small-kernel ConvNets, due to more similar architectures. Our findings are backed up by extensive experiments on both logit-level and feature-level KD "out of the box", with no dedicated architectural nor training recipe modifications. Notably, we obtain the **best-ever pure ConvNet** under 30M parameters with **83.1%** top-1 accuracy on ImageNet, outperforming current SOTA methods including ConvNeXt V2 and Swin V2. We also find that beneficial characteristics of large-kernel ConvNets, e.g., larger effective receptive fields, can be seamlessly transferred to students through this large-to-small kernel distillation. Code is available at: https://github.com/VITA-Group/SLaK.

[1]Department of Mathematics and Computer Science, Eindhoven University of Technology [2]Department of Electrical and Computer Engineering, University of Texas at Austin [3]JD Explore Academy [4]Department of Computer Science, University of Liverpool. Correspondence to: Tianjin Huang <t.huang@tue.nl>, Shiwei Liu <shiwei.liu@austin.utexas.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

1. Introduction

Transformers (Vaswani et al., 2017) have shined in the past two years, bringing a historical revolution in artificial intelligence, from the foundation models in natural language processing (Brown et al., 2020; Ramesh et al., 2022; Du et al., 2022; Li et al., 2023; Chen et al., 2023; Chowdhery et al., 2022) to the Vision Transformers in computer vision (Dosovitskiy et al., 2021; Liu et al., 2021), to biological sciences (Jumper et al., 2021), etc. By leveraging larger and larger models and labeled data, Transformers continuously establish new state-of-the-art (SOTA) performance bars in various fields. Despite their impressive performance, Transformers arguably cause a prohibitive competition of gigantic models in academia and industry, pushing the SOTA model size beyond the reach of common hardware.

On the other hand, Convolutional Neural Networks (ConvNets) still hold their merits in computer vision scenarios with limited computational resources, such as Edge AI (Li et al., 2019) and artificial intelligence of things (AIoT) (Zhang & Tao, 2020). Compared to advanced Vision Transformers, small-kernel ConvNets like ResNets (He et al., 2016b) and MobileNets (Howard et al., 2017) are generally less expensive to train and infer, despite typically having lower performance. Therefore, improving the performance of small-kernel ConvNets to the state-of-the-art levels achieved by advanced Vision Transformers while maintaining model size affordable has excellent value in real-world environments. However, this is a non-trivial research question since previous work (Yao et al., 2023) has shown that distilling knowledge from Vision Transformers to small-kernel ConvNets can be ineffective or even detrimental. In this study, we carry out a first-of-its-kind study unveiling that modern large-kernel ConvNets, a compelling competitor to Vision Transformers, are more effective, and easy-to-adopt teachers for small-kernel ConvNets due to their architectural similarities.

Knowledge Distillation (KD) (Hinton et al., 2015) with its variants (Yuan et al., 2020; Zhou et al., 2021; Zhao et al., 2022) has evolved as one of the leading approaches to enhancing the performance of deep neural networks. The basic concept is to utilize larger teacher models to generate pseudo-labels for smaller models, which are subsequently

trained to mimic the teacher's behavior, without increasing the model size. While distilling knowledge from advanced Vision Transformers to small-kernel ConvNets seems to be a logical choice to boost the performance of the latter, doing so has little efficacy likely due to the inherent architectural discrepancy in between (Yao et al., 2023).

Recently emerged **large-kernel ConvNets**, e.g., ConvNeXt (Liu et al., 2022c) and RepLKNet (Ding et al., 2022), demonstrate that pure convolutional models can deliver the same excellence as Vision Transformers when equipped with similar advanced designs, whose performance can be further stretched out by purely enlarging kernel size to 51×51 in SLaK (Liu et al., 2022a). Yet, the effectiveness of large kernel ConvNets in teaching compact models remains to lack intuition and is unexplored.

Intuitively, large-kernel ConvNets have three advantages (which can be satisfied simultaneously or partially) compared to Vision Transformers, as teachers for small-kernel ConvNets: (1) equally good accuracy; (2) similar or even larger effective receptive field (ERF); (3) more importantly, convolutional operations instead of self-attention modules. However, these modern ConvNets (e.g., ConvNexT and SLaK) also share many discrepancies with conventional small-kernel ConvNets, including but not limited to "patchify stem", inverted bottleneck, BatchNorm instead of LayerNorm, and replacing ReLU with GELU (Liu et al., 2022c). Hence it remains unclear whether large-kernel ConvNets can perform better than Vision Transformers or not when distilled to small-kernel ConvNets.

In this paper, we conduct a systematic comparison between modern large-kernel ConvNets (such as ConvNeXt and SLaK) and advanced Vision Transformers (such as ViT (Dosovitskiy et al., 2021), Swin (Liu et al., 2021), and CSWin (Dong et al., 2022)) as teacher models, when distilled into small-kernel ConvNet student. We discover that large-kernel ConvNets are significantly more effective teachers than Vision Transformers for small-kernel ConvNets, for both feature-level and logit-level KD approaches. We also find that our distilled models enjoy larger ERF and better robustness than others, indicating that, besides accuracy, other good properties of large kernels can also be seamlessly transferred to small kernels through our large-to-small kernel distillation paradigm. Our contribution can be summarized as follows:

- We conduct a pioneering empirical study of large-to-small kernel distillation. Through a thorough comparison of advanced Vision Transformers (including ViT, Swin, and CSWin) and modern large-kernel ConvNets (ConvNeXt and SLaK) when distilled to small-kernel ConvNets, we discover two principles for ConvNet distillation: ❶ large-kernel ConvNets function as more effective teachers than Transformers for small-kernel ConvNets; ❷ among large-kernel teachers, students obtain greater benefits from larger kernels compared with the smaller ones.

- Following the above principles without any dedicated architectural and training designs, we favorably distill the recently proposed 51×51 SLaK-T into 7×7 ConvNeXt-T, obtaining the best-ever 30M ConvNet with **83.1%** top-1 accuracy on ImageNet, outperforming the current state-of-the-arts ConvNeXt V2 and Swin V2. Interestingly, the distilled ConvNeXt-T even outperforms the strong result of its teacher by 0.6%.

- More interestingly, we found that students distilled from larger kernels are automatically embedded with better robustness, as well as larger and denser ERF than the ones distilled from small kernels or Transformers. Given the recently emerging trend that correlates larger ERF with better performance (Ding et al., 2022; Kim et al., 2021; Liu et al., 2022a; Dai et al., 2022; Yang et al., 2022a), this observation also explains the success of large-to-small kernel distillation.

2. Related Work

2.1. Knowledge Distillation

Knowledge distillation (KD) is proposed by Hinton et al. (2015), to transfer knowledge from one teacher model to another student model, by forcing the latter to mimic the prediction of the teacher. Since then, various variants of knowledge distillation have been proposed (Fang et al., 2022; 2021; Xue et al., 2021), which can be roughly categorized into two groups: 1) Distillation from logits (Cho & Hariharan, 2019; Yang et al., 2019; Mirzadeh et al., 2020; Yang et al., 2022b; Yao et al., 2023). Logit-level distillation mainly try to minimize the KL-Divergence between prediction logits of teachers and students. Recently, Zhao et al. (2022) decouple the classical KD loss into two parts, i.e., target class knowledge distillation and non-target class knowledge distillation, achieving better results. Yang et al. (2022b) revise the formulation of classical KD and found that besides cross-entropy (CE) loss, it also contains an extra loss which contains the knowledge of all classes except the target class. By adding teachers' target output as an extra soft loss, they can outperform the previous logit distillation approaches, like KD, DKD (Yang et al., 2022c), etc. 2) Distillation from representation (Heo et al., 2019a;b; Yang et al., 2021; Huang & Wang, 2017; Kim et al., 2018; Park et al., 2019; Yim et al., 2017; Zagoruyko & Komodakis, 2016; Wei et al., 2022). Romero et al. (2014) distill the learned knowledge from the intermediate feature directly. Park et al. (2019) extract relations from the feature map and transfer the relation instead. Chen et al. (2021) distill knowledge from multi-level features map.

Recently, Yang et al. (2022d) proposes to let the student model generate the teacher model's feature instead of mimicking (MGD), improving the effectiveness of feature distillation. Yang et al. (2022e) further comprehensively studies the effect of mimicking KD and Generation KD on ViT distillation, resulting in a suite of principles for feature-based ViT distillation.

Li et al. (2022) introduce spatial-channel token distillation to mix up the information in both the spatial and channel dimensions for Vision MLP. Multiple receptive tokens are applied in dense prediction tasks to indicate the pixels of interest in the feature map, with a distillation mask generated via pixel-wise attention (Huang et al., 2022b). Huang et al. (2022a) revisit the fact the training with much stronger teachers could significantly hurt student's performance (Cho & Hariharan, 2019) and address it by simply preserving the relations between the prediction of teacher and student with a correlation-based loss. Very recent work (Yao et al., 2023) unveils that distilling knowledge (e.g., global information) from Vision Transformers to ConvNets is ineffective likely due to their architectural gaps. In this work, instead of distilling knowledge from Transformers, we introduce large-to-small kernel distillation, which can seamlessly transfer preferable knowledge (e.g., accuracy, global information, and robustness) to small-kernel ConvNets.

2.2. Large Kernel Convolutions

Large kernel convolutions date back to the 2010s, where AlexNet (Krizhevsky et al., 2012) adopts 11×11 kernels in the first convolutional layer and Inception (Szegedy et al., 2015; 2017) stacks a combination of 1×7 and 7×1 kernels. Global Convolutional Network (Peng et al., 2017) constructs ConvNets with a pair of stacked large convolution, with kernel size up to 25×1 + 1×25. Since the popularity of VGG (Simonyan & Zisserman, 2014), people favor small stacked 3×3 kernels to build ConvNets (He et al., 2016b; Howard et al., 2017; Xie et al., 2017; Huang et al., 2017). Although computationally efficient, it is not very efficient for small kernels to achieve large effective receptive fields (ERF), even with over 100 layers in the model.

Inspired by the local window (at least 7×7) self-attention used in Swin Transformers (Liu et al., 2021), Liu et al. (2022c) revisit the use of large kernel-sized convolutions for ConvNets, finding that 7×7 depthwise convolutions consistently outperform 3×3 in ConvNeXt. RepLKNet (Ding et al., 2022) continues to scale kernel size to 31×31 using Structural Reparameterization while achieving comparable or superior results than Swin Transformer. SLaK (Liu et al., 2022a) pushes along the direction of pure ConvNets with extremely large kernels up to 51×51, which has never been discussed before. More recently, Chen et al. (2022); Xiao et al. (2022) reveals the feasibility of large kernels for 3D ConvNets and time series classification, respectively. SegNeXt (Guo et al., 2022) ensembles a novel convolutional attention network with multiple large kernels, showing strong results on semantic segmentation. Our paper differs from the previous work and focuses on an empirical pilot study of knowledge distillation from large kernels to small kernels.

3. Experimental Setup

In this section, we describe the experimental setup and benchmarks used. Our goal is to comprehensively compare Vision Transformers and modern large-kernel ConvNets in the context of knowledge distillation and to study which is more suitable as teachers for small-kernel ConvNets. To enable fair comparisons, we carefully investigate several key components and summarize them below.

Evaluation Metrics. Given a teacher model (T) with high accuracy on a task $acc\,(teacher)$ and a student model (S) with lower accuracy $acc\,(student)$, we can improve the accuracy of the latter to $acc\,(distilled)$ by via knowledge distillation. To enable comparisons among different teacher models, we choose two metrics to report, Direct Gain and Effective Gain. Direct Gain refers to the direct performance difference with and without knowledge distillation, i.e.,

$$\text{Direct Gain} = acc\,(distilled) - acc\,(student) \quad (1)$$

Ideally, we hope all the teachers have the same accuracy to make a fair comparison solely for their distillation effectiveness. However, different models inevitably have accuracy discrepancies, impacting our comparisons. To mitigate this undesirable effect, we scale the Direct Gain by teachers' accuracy, resulting in Effective Gain:

$$\text{Effective Gain} = \frac{acc\,(distilled) - acc\,(student)}{acc\,(teacher)} \quad (2)$$

In all our experiments, we report these two metrics.

Dataset, Teacher and Student Models. We conduct experiments on the commonly used ImageNet-1K dataset (Russakovsky et al., 2015) containing 1k classes, 1,281,167 training images, and 50,000 validation images.

We have two main distillation pipelines: Pipeline I: Large-kernel ConvNets to small-kernel ConvNets distillation; Pipeline II: Transformers to small-kernel ConvNets distillation. For both pipelines, we opt for two ConvNets as our students: ResNet-50 with 3×3 kernels as the most representative conventional ConvNet, and ConvNeXt-T[1] with 7×7 kernels as the most representative modern ConvNet. ConvNeXt-T also serves suitable small-kernel ConvNet when it comes to teachers like SLaK-T (51×51 kernels), ViT (global attention), Swin-T (at least 7×7 windows), and CSWin-T (global cross-shaped windows). Our

[1]Follow (Liu et al., 2022a), we add a BatchNorm layer after 7×7 kernels since it gives us consistently better accuracy.

teacher models are ConvNeXt-T and SLaK for Pipeline I; ViT-S, Swin-T, and CSWin-T for Pipeline II. Overall, the teachers in Pipeline I and Pipeline II are of similar sizes and comparable accuracy, ensuring a fair comparison. The pre-trained weights of SLaK-T, ConvNeXt-T, and CSWin-T are downloaded from their official GitHub repositories. The pre-trained weights of ViT-S and Swin-T are downloaded from TIMM (Wightman, 2019).

Training Recipe. In our study, we use a set of training recipes closely following DeiT (Touvron et al., 2021b), Swin Transformers (Liu et al., 2021), and ConvNeXt (Liu et al., 2022c). We use AdamW optimizer (Loshchilov & Hutter, 2019) and train models for 120 epochs (Section 4.1) and 300 epochs (Section 4.2) with a batch size of 4096, and a weight decay of 0.05. The learning rate is 4e-3 with a 20-epoch linear warmup followed by a cosine decaying schedule.

For data augmentation, we use the default RandAugment (Cubuk et al., 2020) in Timm (Wightman, 2019) – "rand-m9-mstd0.5-inc1", Label Smoothing (Szegedy et al., 2016) coefficient of 0.1, Mixup (Zhang et al., 2017) with $\alpha = 0.8$, Cutmix (Yun et al., 2019) with $\alpha = 1.0$, Random Erasing (Zhong et al., 2020) with $p = 0.25$, Stochastic Depth with a drop rate of 0.1 for ConvNeXt-T and 0.0 for ResNet-50, and Layer Scale (Touvron et al., 2021c) of the initial value of 1e-6. We train all models with 4 NVIDIA A100 GPUs.

Distillation Methods. To draw solid conclusions, we adopt both logit-level distillation and feature-level distillation in this study. Without loss of generality, we opt for the widely-used Knowledge Distillation (KD) (Hinton et al., 2015) as our logit-level method. Additionally, the recently proposed New Knowledge Distillation (NKD) (Yang et al., 2022b) is also adopted given its strong results over KD and DKD (Yang et al., 2022c). We adopt FD (Wei et al., 2022) as our feature-level distillation due to its superior performance. For clarity, we provide the following formal definitions for each method. Let Z_t, Z_s be the logits of the teacher model and the student model, respectively; $KL(\cdot)$, $\mathcal{L}_{CE}(\cdot)$, and $\phi(\cdot)$ is the Kullback-Leiler divergence loss, the cross-entropy loss, and the softmax function, respectively. We denote τ as the temperature hyperparameter of KD, C as the classes, (x,y) as the inputs, and λ as the coefficient.

• *Knowledge Distillation (KD)* (Hinton et al., 2015). It consists of the KL divergence loss and the cross-entropy loss. Concretely, the cross-entropy loss encourages the student model to learn knowledge from the true label and the KL divergence transfers the knowledge of the teacher to the student. Formally, it is expressed as follows:

$$\mathcal{L}_{KD} = (1-\lambda)\mathcal{L}_{CE}(\phi(Z_s), y) + \lambda\tau^2 KL(\phi(Z_s/\tau), \phi(Z_t/\tau)) \tag{3}$$

• *New Knowledge Distillation (NKD)* (Yang et al., 2022b). Inspired by the cross-entropy loss where the sum of the two distributions are equal, NKD normalizes both the non-target student and teacher output probability such that they meet this constraint as well. Besides, NKD also proposes the soft loss that regards the teacher's target output as the soft target directly. Therefore, NKD consists of the original loss, the non-target distributed loss, and the target soft loss. Formally it is expressed as follows:

$$\begin{aligned}\mathcal{L}_{NKD} = &- log(\phi(Z_s^y)) - \phi(Z_t)^y log(\phi(Z_s)^y) \\ &- \lambda\tau^2 \sum_{i\neq y}^{C} \widehat{T}^i log(\widehat{S}^i)\end{aligned} \tag{4}$$

where $\widehat{T}^i = \frac{\phi(Z_t/\tau)^i}{1-\phi(Z_t/\tau)^y}, \widehat{S}^i = \frac{\phi(Z_s/\tau)^i}{1-\phi(Z_s/\tau)^y}$ refer to the normalized non-target knowledge and the normalized student's non-target output probability respectively.

• *Feature Distillation (FD)* (Wei et al., 2022). Feature distillation distills the learned knowledge from intermediate feature maps. Following (Wei et al., 2022), we apply a 1×1 convolution layer on the top of the student model to align the feature map dimensions of the student model to the teacher model and normalize the output feature map of the teacher model by a whitening operation, implemented by a non-parametric layer normalization operator without scaling and bias. In distillation, we adopt a smooth l_1 loss between the student and teacher feature maps, which is formally expressed as follows:

$$\mathcal{L}_{FD} = \mathcal{L}_{CE}(\phi(Z_s), y) + \tag{5}$$
$$\begin{cases} \sum_{l=1}^{L} \frac{1}{2}(g(S^l) - norm(T^l))^2/\beta, & \|g(S^l) - norm(T^l)\|_1 \leq \beta, \\ \sum_{l=1}^{L} \|g(S^l) - norm(T^l) - 0.5\beta\|_1, & otherwise \end{cases}$$

where L denotes the total number of layers that are used for feature distillation and S^l, T^l denotes the l-th feature map of the student and the teacher, respectively. g is a 1×1 convolution layer and $norm$ is a non-parametric layer normalization for the whitening operation. β is a hyperparameter and is set to 2.0 by default following (Wei et al., 2022).

4. Experimental Results

We report our main results in this section, by evaluating the effectiveness of large-kernel teachers in both logits-level and feature-level KD on ImageNet (Russakovsky et al., 2015).

4.1. Large-Kernel ConvNet vs. Transformer as Teachers

4.1.1. Logit-Level Distillation

We evaluate the performance on logit-level distillation based on two distillation methods: KD (Hinton et al., 2015) and NKD (Yang et al., 2022b). The experiments are conducted using five teachers across convolution-based and transformer-based architectures: ViT-S (Dosovitskiy et al., 2021), ConvNeXt-T (Liu et al., 2022c), Swin-T (Liu et al., 2021), CSWin-T (Dong et al., 2022), SLaK-T (Liu et al.,

Table 1. **NKD (Yang et al., 2022b) results of ResNet-50 and ConvNeXt-T distilled from various teachers on ImageNet.** All models are distilled for 120 epochs. Effective Gain is defined as $\frac{acc\,(distilled)-acc\,(student)}{acc\,(teacher)}$, and Direct Gain is $acc\,(distilled)-acc\,(student)$.

Teacher	Arch. Type	Kernel-Size	Student	Teacher Top-1	Student Top-1	Distilled Top-1	Effective Gain	Direct Gain
ViT-S	Transformer	N/A	ResNet-50	79.8	76.13	77.14	1.20	1.01
Swin-T	Transformer	N/A	ResNet-50	81.3	76.13	77.67	1.89	1.54
CSWin-T	Transformer	N/A	ResNet-50	82.7	76.13	77.68	1.87	1.55
ResNet-50	ConvNet	3×3	ResNet-50	80.4	76.13	77.82	2.1	1.69
ConvNeXt-T	ConvNet	7×7	ResNet-50	82.1	76.13	78.04	2.30	1.91
SLaK-T	ConvNet	51×51	ResNet-50	82.5	76.13	**78.57**	**2.90**	**2.44**
Swin-T	Transformer	N/A	ConvNeXt-T	81.3	81.00	81.10	0.12	0.10
CSWin-T	Transformer	N/A	ConvNeXt-T	82.7	81.00	81.65	0.70	0.65
ResNet-50	ConvNet	3×3	ConvNeXt-T	80.4	81.00	81.15	0.19	0.15
ConvNeXt-T	ConvNet	7×7	ConvNeXt-T	82.1	81.00	81.77	0.93	0.77
SLaK-T	ConvNet	51×51	ConvNeXt-T	82.5	81.00	**82.17**	**1.42**	**1.17**

Table 2. **KD (Hinton et al., 2015) results of ResNet-50 and ConvNeXt-T distilled from various teachers on ImageNet.** All models are distilled for 120 epochs. Effective Gain is defined as $\frac{acc\,(distilled)-acc\,(student)}{acc\,(teacher)}$, and Direct Gain is $acc\,(distilled)-acc\,(student)$.

Teacher	Arch. Type	Kernel-Size	Student	Teacher Top-1	Student Top-1	Distilled Top-1	Effective Gain	Direct Gain
ViT-S	Transformer	N/A	ResNet-50	79.8	76.13	76.96	1.04	0.83
Swin-T	Transformer	N/A	ResNet-50	81.3	76.13	76.87	0.91	0.74
CSWin-T	Transformer	N/A	ResNet-50	82.7	76.13	76.77	0.77	0.64
ConvNeXt-T	ConvNet	7×7	ResNet-50	82.1	76.13	76.93	0.97	0.80
SLaK-T	ConvNet	51×51	ResNet-50	82.5	76.13	**77.05**	**1.12**	**0.92**
Swin-T	Transformer	N/A	ConvNeXt-T	81.3	81.00	80.93	-0.08	-0.07
CSWin-T	Transformer	N/A	ConvNeXt-T	82.7	81.00	81.18	0.22	0.18
ConvNeXt-T	ConvNet	7×7	ConvNeXt-T	82.1	81.00	81.64	0.77	0.64
SLaK-T	ConvNet	51×51	ConvNeXt-T	82.5	81.00	**81.86**	**1.04**	**0.86**

2022a), and two student models: ConvNeXt-T (Liu et al., 2022c) and ResNet-50 (He et al., 2016a). We exclude ViT-S as a teacher for ConvNeXt-T since the performance of the former is worse than the latter. We follow (Liu et al., 2022a) and train all models for 120 epochs to simply show the performance trend. Later on in Section 4.2, we will adopt the full training recipe and train our models for 300 epochs, to enable fair comparisons with state-of-the-art models. Consequently, "Student Top-1" is also obtained by 120-epoch training. We sweep over temperatures $\{1, 2, 5, 10, 20\}$ for all approaches. We summarize our main observations below:

❶ Large-kernel ConvNets serve as better teachers than Transformers for small-kernel ConvNets. Table 1 and Table 2 show the results of NKD and KD across various teachers, respectively. Overall, we can obverse that large-kernel teachers such as SLaK-T and ConvNeXt-T outperform all Transformer teachers by a good margin in terms of both Effective Gain and Direct Gain metrics. While the teacher accuracy of ConvNeXt falls short of the best Transformers CSWin-T by 0.6%, its student models consistently outperform the latter by up to 0.46% Direct Gain and 0.55% Effective Gain. Moreover, Direct Gain of SLaK-T reaches 2.44% and 1.17% for ResNet-50 and ConvNeXt-T, respectively, which is notably higher than the best results achieved by Transformer teachers, i.e., 1.55% and 0.65%.

❷ Students benefit more from larger kernels. As teacher models, SLaK-T increases the kernel size of ConvNeXt-T from 7×7 to 51×51 via sparsity and factorization, outperforming the latter by 0.4%. Among large-kernel teachers, the student models distilled from SLaK-T consistently outperform those distilled from ConvNeXt-T under both Effective Gain and Direct Gain metrics. It indicates that the benefits of larger kernels over small kernels can be effectively distilled into student models and larger kernels teach better than the small kernels. The benefits of large-kernel ConvNets over Vision Transformers can also be generalized to lightweight ConvNets (please refer to Appendix D).

❸ Large-kernel ConvNets help student train faster. It takes 300 epochs for ConvNeXt-T to reach 82.1% accuracy on ImageNet under supervised training (Liu et al., 2022c). It

Table 3. **Feature distillation results of ResNet-50 distilled from various teacher models on ImageNet dataset.** The hyper-parameter L is set to 1. Effective Gain is defined as $\frac{acc\,(distilled)-acc\,(student)}{acc\,(teacher)}$, and Direct Gain is $acc\,(distilled) - acc\,(student)$.

Teacher	Arch. Type	Student	Teacher Top-1	Student Top-1	Distilled Top-1	Effective Gain	Direct Gain
Feature Distillation							
ViT-S	Transformer	ResNet-50	79.8	76.13	76.81	0.85	0.68
Swin-T	Transformer	ResNet-50	81.3	76.13	76.77	0.78	0.64
ConvNeXt-T	ConvNet	ResNet-50	82.1	76.13	76.73	0.61	0.70
SLaK-T	ConvNet	ResNet-50	82.5	76.13	**76.85**	**0.87**	**0.72**
Feature + Logits (NKD) Distillation							
ViT-S	Transformer	ResNet-50	79.8	76.13	76.84	0.89	0.71
Swin-T	Transformer	ResNet-50	81.3	76.13	77.33	1.47	1.20
ConvNeXt-T	ConvNet	ResNet-50	82.1	76.13	77.85	2.09	1.72
SLaK-T	ConvNet	ResNet-50	82.5	76.13	**77.99**	**2.25**	**1.86**

is worth noting that our distilled ConvNeXt-T, when distilled from the 51×51 kernel SLaK-T, reaches the performance of its 300-epoch supervised performance in only 120 epochs, highlighting the benefits of large-to-small kernel distillation.

Table 4. **Feature distillation with varying the number of layers L. Results of ResNet-50 distilled from SLaK-T and Swin-T.** Direct Gain is defined as $acc\,(distilled) - acc\,(student)$.

L	No Distillation	FD Distillation			
	ResNet-50	Teacher:Swin-T		Teacher:SLaK-T	
	Top-1	Top-1	Direct Gain	Top-1	Direct Gain
1	76.13	76.77	0.64	76.85	0.72
2		76.76	0.63	76.79	0.66
3		76.81	0.68	76.94	0.81
4		76.73	0.60	76.92	0.79

4.1.2. Feature-Level Distillation

Besides only matching the logits, many works aim to minimize the distance of intermediate features between teachers and students (Heo et al., 2019a;b; Yang et al., 2021; Huang & Wang, 2017; Kim et al., 2018). To comprehensively understand the effect of Transformers and large-kernel ConvNets based teachers, we also conduct comparisons on feature-level distillation. Specifically, we follow the feature-level KD (Wei et al., 2022) that chooses the output of the last stage by default to transfer knowledge. Moreover, it is common to combine logit distillation with feature distillation to further improve performance (Yang et al., 2022e; 2021). Here, we combine FD (Wei et al., 2022) with NKD and evaluate different teacher models with a student model ResNet-50. Concretely, we combine feature-level and logit-level distillation by minimizing the FD loss and NKD loss together. In addition, we also conduct comparisons on multi-layer feature distillation. With $L = i$, it denotes that the outputs of the last i number of stages are used to construct the feature distillation loss. For example, with $L = 2$, the outputs of the last two stages are used for feature distillation. Results are shown in Table 3 and Table 4.

In Table 3, we observe that large-kernel ConvNets again consistently produce more performant student models than transformer-based teachers for feature distillation. Interestingly, while SLaK-T only surpasses Swin-T by 0.08% Direct Gain in feature distillation, the corresponding number increases to 0.66% when incorporated with logit distillation. This result highlights the benefits of large-kernel ConvNets in combining logits and features. In Table 4, we observe that the student model distilled from SLaK-T outperforms those distilled from Swin-T with varying L, indicating the superiority of large-kernel teachers over transformer-based teachers is also held when using FD with feature maps of multiple layers, i.e. $L > 1$.

4.2. Scaling to Longer Training

Recent study (Beyer et al., 2022) have shown that knowledge distillation requires an atypically larger number of training epochs to reach the best performance, much more than commonly used in supervised learning. Here, we also extend the training time from 120 to 300 epochs and report the performance of ResNet-50 distilled from both large-kernel teachers and transformer-based teachers. The results are shown in Table 5.

We observe that the performance of ResNet-50 receives a 2.05% performance boost when extending training epochs from 120 epochs to 300 epochs, which is in line with the observation in (Beyer et al., 2022). It is clear to see that the performance trend of longer training schedules is closely consistent with the short schedules. The best student model is achieved by the SLaK-T teacher among all five teachers, indicating that the benefits of large-kernel teachers over transformers-based teachers also held with longer training.

Best-Ever 30M ConvNet. Based on the above observations,

Table 5. **NKD (Yang et al., 2022b) results of ResNet-50 distilled from various teacher models on ImageNet dataset.** All models are distilled for 300 epochs. Effective Gain is defined as $\frac{acc\,(distilled) - acc\,(student)}{acc\,(teacher)}$, and Direct Gain is $acc\,(distilled) - acc\,(student)$.

Teacher	Arch. Type	Student	Teacher Top-1	Student Top-1	Distilled Top-1	Effective Gain	Direct Gain
ViT-S	Transformer	ResNet-50	79.8	78.76	78.88	0.15	0.12
Swin-T	Transformer	ResNet-50	81.3	78.76	79.44	1.45	1.18
ConvNeXt-T	ConvNet	ResNet-50	82.1	78.76	80.09	1.6	1.33
SLaK-T	ConvNet	ResNet-50	82.5	78.76	**80.24**	**1.8**	**1.48**

Table 6. **Top-1 accuracy of various SOTA models on ImageNet-1K.**

Model	Method	#Training Epochs	Image Size	#Param.	FLOPs	Top-1 Acc
ResNet-50 (He et al., 2016b)	Supervised	90	224×224	26M	4.1G	76.5
ResNeXt-50-32×4d (Xie et al., 2017)	Supervised	90	224×224	25M	4.3G	77.6
ResMLP-24 (Touvron et al., 2021a)	Supervised	400	224×224	30M	6.0G	79.4
DeiT-S (Touvron et al., 2021b)	Supervised	300	224×224	22M	4.6G	79.8
Swin-T (Liu et al., 2021)	Supervised	300	224×224	28M	4.5G	81.3
TNT-S (Han et al., 2021)	Supervised	300	224×224	24M	5.2G	81.3
T2T-ViT$_t$-14 (Yuan et al., 2021)	Supervised	310	224×224	22M	6.1G	81.7
ConvNeXt V1-T (Liu et al., 2022c)	Supervised	300	224×224	29M	4.5G	82.1
SLaK-T (Liu et al., 2022a)	Supervised	300	224×224	30M	5.0G	82.5
Swin V2-T (Liu et al., 2022b)	Self-Supervised	300	256×256	28M	6.6G	82.8
ConvNeXt V2-T (Woo et al., 2023)	Self-Supervised	900	224×224	29M	4.5G	83.0
ResNet-50 (Beyer et al., 2022)	Knowledge Distillation	9600	224×224	26M	4.1G	82.8
ConvNeXt L2S-T	Knowledge Distillation	300	224×224	29M	4.5G	**83.1**
Swin-S (Liu et al., 2021)	Supervised	300	224×224	50M	8.7G	83.0
ConvNeXt V1-S (Liu et al., 2022c)	Supervised	300	224×224	50M	8.7G	83.1
SLaK-S (Liu et al., 2022a)	Supervised	300	224×224	55M	9.8G	83.8
Swin V2-S (Liu et al., 2022b)	Self-Supervised	300	256×256	50M	12.6G	84.1
ConvNeXt L2S-S	Knowledge Distillation	300	224×224	50M	8.7G	**84.2**

we directly distill 51×51 SLaK-T into a 7×7 ConvNeXt-T; SLaK-S into ConvNeXt-S, via NKD for 300 epochs, dubbed **ConvNeXt L2S-T/S**. ConvNeXt L2S-T achieves a new state-of-the-art **top-1 accuracy of 83.1% for pure ConvNets on ImageNet**, outperforming its supervised counterpart by 1.0% accuracy. It is also very encouraging to observe that our distilled ConvNeXt-T, with no dedicated architectural designs, is able to slightly outperform the current state-of-the-art ConvNeXt V2-T (Woo et al., 2023), which is trained by MAE-based self-supervised learning with an improved architecture for 900 epochs in total. Our model also achieves better accuracy than the best ResNet-50 model in the literature (Beyer et al., 2022) which is trained with knowledge distillation for 9600 epochs.

5. What Else are Transferrable from Larger Kernels Teachers?

We then go beyond the accuracy and evaluate other important properties of students distilled from different teachers, by visualizing the effective receptive field (ERF) and testing the robustness on several ImageNet-level benchmarks.

5.1. Transferring Effective Receptive Fields (ERF)

The concept of the effective receptive field (Araujo et al., 2019; Luo et al., 2016) is an important concept in computer vision. For the output unit of one neural network layer, ERF is defined as the region containing any input pixel with a non-negligible impact on that unit (Araujo et al., 2019). Intuitively, anywhere in an input image outside the receptive field of a unit does not affect its output value. It is generally accepted that both large-kernel ConvNets and Vision Transformers have larger ERF, which in turn helps them outperform traditional small-kernel models. While we have known that distillation brings significant performance improvement to the student model, it remains very interesting to investigate whether desirable characteristics like large ERF can be distilled into small-kernel ConvNets. To do this, we visualize the ERFs of students that are distilled from large-kernel ConvNets and Vision Transformers.

Following (Ding et al., 2021; Liu et al., 2022a), we sample and resize 50 images from the validation set to 1024×1024, and measure the contribution of each pixel on input images to the central point of the feature map generated in the last layer. The contribution scores are further accumulated and projected to a 1024×1024 matrix. The visualization is shown in Figure 1. We find that students distilled from SLaK-T are automatically embedded with larger and denser ERF than the ones from Swin-T and CSWin-T, albeit all of these teacher models achieve sufficient large ERF with global or large self-attention. This result also helps us better understand the success of large kernels in ConvNet distil-

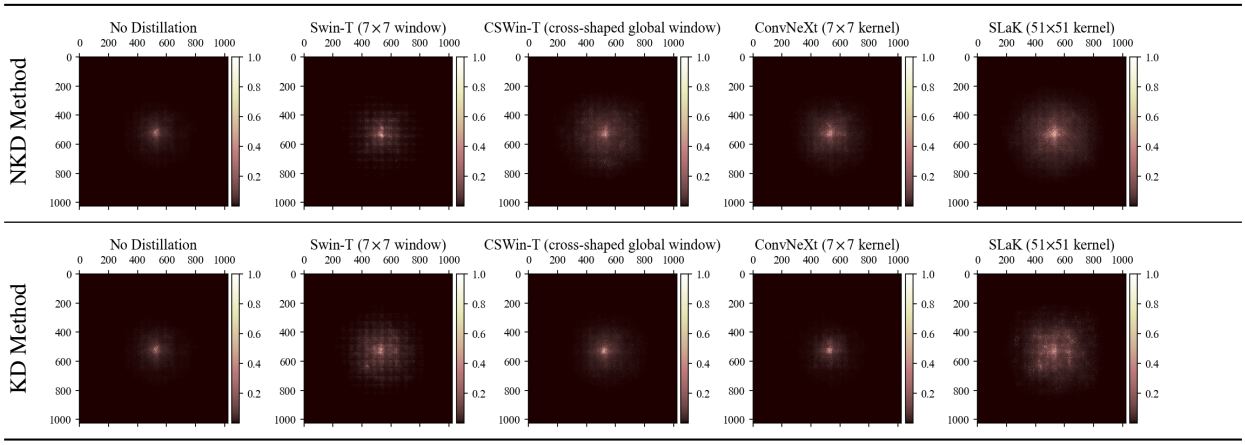

Figure 1. **Effective receptive field (ERF) of the ConvNeXt-T distilled from various teachers**. Our student model is ConvNeXt-T with 7×7 kernels. The left figures refer to the supervised ConvNeXt-T without distillation, and the rest figures are from distilled ConvNeXt-T. Overall, the students distilled from 51×51 SLaK have larger and denser ERF than the students from Transformer teachers.

lation, that is, large kernels are more effective at distilling large REF to small kernels, improving the inherent deficiency of the latter.

5.2. Transferring Robustness

Recent studies on out-of-distribution robustness (Wang et al., 2022; Bai et al., 2021) show that transformers and the large-kernel ConvNets embrace more robustness than the ConvNets with small-kernel. However, it is seldom explored whether such superiority in robustness can be transferred to small-kernel models through knowledge distillation. To answer this question, we directly test the different student models on several robustness benchmarks including ImageNet-R (Hendrycks et al., 2021a), ImageNet-A (Hendrycks et al., 2021b), ImageNet-Sketch (Wang et al., 2019), and ImageNet-C (Hendrycks & Dietterich, 2019) datasets. Mean corruption error (mCE) is reported for ImageNet-C and top-1 accuracy is reported for the rest.

The results are shown in Table 7. We can clearly see that the students distilled from modern ConvNets exhibit very promising robustness, consistently outperforming the students that are learned from state-of-the-art Transformers. Among large-kernel teachers, SLaK-T transfers better robustness to students than ConvNeXt, even though with lower robustness as teachers. However, robust Transformers do not necessarily transfer to small kernel students. For example, ResNet-50 models distilled from Swin-T and ViT-S even decrease their top-1 accuracy on ImageNet-SK/A datasets. This delivers a strong signal that large kernels are stronger teachers than advanced Vision Transformers and small kernels in terms of in-distribution and out-of-distribution.

More comparison with state-of-the-art models can be referred to Appendix B, in which the distilled ConvNeXt student model surpasses the performance of most baseline models and attains a level of robustness comparable to the large-kernel-based model such as SLaK-Tiny, as well as high-capacity models like Swin-B (Liu et al., 2021) and RVT-B (Mao et al., 2022).

6. Conclusions

We have explored distilling recently popular large-kernel ConvNets into small-kernel ConvNets. This empirical study is meaningful due to the fact that small-kernel ConvNets remain valuable for practical application with limited resources. Our paper comprehensively compares SOTA large-kernel ConvNets (ConvNeXt and SLaK) with various state-of-the-art Vision Transformers (such as ViT, Swin, and CSWin) in several KD scenarios, including logit distillation, feature distillation, transferring of ERF, and transferring of robustness. Our results demonstrate that large-kernel ConvNets are all-around stronger teachers than Vision Transformers for transferring knowledge to small-kernel ConvNets, and suggest that the merits of convolutions are not easily fading away. We hope our empirical study will further encourage the community to revisit ConvNets.

7. Acknowledgement

S. Liu and Z. Wang are in part supported by the NSF AI Institute for Foundations of Machine Learning (IFML). Part of this work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. NWO2021.060, EINF-2694 and EINF-2943/L1.

References

Araujo, A., Norris, W., and Sim, J. Computing receptive fields of convolutional neural networks. *Distill*, 4(11): e21, 2019.

Table 7. **Robustness evaluation of distilled students.**

Teachers	Arch. Type	Student	Clean Accuracy ↑	C ↓	SK ↑	R ↑	A ↑
Robustness of Teacher Models							
ViT-S	Transformer	-	79.8	47.15	26.92	39.88	12.13
Swin-T	Transformer	-	81.3	46.07	29.05	41.20	21.17
CSWin-T	Transformer	-	82.7	**39.47**	33.52	44.99	**31.58**
ConvNeXt-T	ConvNet	-	82.1	41.99	**33.85**	**47.13**	24.02
SLaK-T	ConvNet	-	82.5	41.17	32.41	45.33	29.89
Robustness of ResNet-50 Students							
-	-	ResNet-50	76.13	54.94	26.94	35.42	4.68
ViT-S	Transformer	ResNet-50	77.14	53.21	25.89	38.91	3.72
Swin-T	Transformer	ResNet-50	77.67	53.46	26.07	37.59	5.45
CSWin-T	Transformer	ResNet-50	77.68	53.24	27.00	38.43	6.27
ConvNeXt-T	ConvNet	ResNet-50	78.04	52.43	27.74	39.02	5.91
SLaK-T	ConvNet	ResNet-50	78.57	**52.09**	**28.54**	**40.16**	**7.12**
Robustness of ConvNeXt-T Students							
-	-	ConvNeXt-T	81.00	44.69	32.62	45.72	20.04
Swin-T	Transformer	ConvNeXt-T	81.10	44.97	31.24	43.35	17.89
CSWin-T	Transformer	ConvNeXt-T	81.65	42.69	33.34	45.19	21.19
ConvNeXt-T	ConvNet	ConvNeXt-T	81.77	43.17	33.82	46.49	21.06
SLaK-T	ConvNet	ConvNeXt-T	82.17	**42.02**	**35.13**	**47.50**	**24.27**

Bai, Y., Mei, J., Yuille, A. L., and Xie, C. Are transformers more robust than cnns? *Advances in Neural Information Processing Systems*, 34:26831–26843, 2021.

Beyer, L., Zhai, X., Royer, A., Markeeva, L., Anil, R., and Kolesnikov, A. Knowledge distillation: A good teacher is patient and consistent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10925–10934, 2022.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Chen, P., Liu, S., Zhao, H., and Jia, J. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5008–5017, 2021.

Chen, T., Zhang, Z., Jaiswal, A., Liu, S., and Wang, Z. Sparse moe as the new dropout: Scaling dense and self-slimmable transformers. *arXiv preprint arXiv:2303.01610*, 2023.

Chen, Y., Liu, J., Qi, X., Zhang, X., Sun, J., and Jia, J. Scaling up kernels in 3d cnns. *arXiv preprint arXiv:2206.10555*, 2022.

Cho, J. H. and Hariharan, B. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4794–4802, 2019.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 702–703, 2020.

Dai, J., Shi, M., Wang, W., Wu, S., Xing, L., Wang, W., Zhu, X., Lu, L., Zhou, J., Wang, X., et al. Demystify transformers & convolutions in modern image deep networks. *arXiv preprint arXiv:2211.05781*, 2022.

Ding, X., Chen, H., Zhang, X., Han, J., and Ding, G. Repmlpnet: Hierarchical vision mlp with re-parameterized locality. *arXiv preprint arXiv:2112.11081*, 2021.

Ding, X., Zhang, X., Zhou, Y., Han, J., Ding, G., and Sun, J. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. *arXiv preprint arXiv:2203.06717*, 2022.

Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., Chen, D., and Guo, B. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12124–12134, 2022.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer,

M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Du, N., Huang, Y., Dai, A. M., Tong, S., Lepikhin, D., Xu, Y., Krikun, M., Zhou, Y., Yu, A. W., Firat, O., et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pp. 5547–5569. PMLR, 2022.

d'Ascoli, S., Touvron, H., Leavitt, M. L., Morcos, A. S., Biroli, G., and Sagun, L. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, pp. 2286–2296. PMLR, 2021.

Fang, G., Bao, Y., Song, J., Wang, X., Xie, D., Shen, C., and Song, M. Mosaicking to distill: Knowledge distillation from out-of-domain data. *Advances in Neural Information Processing Systems*, 34:11920–11932, 2021.

Fang, G., Mo, K., Wang, X., Song, J., Bei, S., Zhang, H., and Song, M. Up to 100x faster data-free knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6597–6604, 2022.

Guo, M.-H., Lu, C.-Z., Hou, Q., Liu, Z., Cheng, M.-M., and Hu, S.-M. Segnext: Rethinking convolutional attention design for semantic segmentation. *arXiv preprint arXiv:2209.08575*, 2022.

Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., and Wang, Y. Transformer in transformer. *Advances in Neural Information Processing Systems*, 34, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016a.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016b. doi: 10.1109/CVPR.2016.90.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021a.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021b.

Heo, B., Kim, J., Yun, S., Park, H., Kwak, N., and Choi, J. Y. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1921–1930, 2019a.

Heo, B., Lee, M., Yun, S., and Choi, J. Y. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3779–3787, 2019b.

Hinton, G., Vinyals, O., Dean, J., et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Huang, T., You, S., Wang, F., Qian, C., and Xu, C. Knowledge distillation from a stronger teacher. *arXiv preprint arXiv:2205.10536*, 2022a.

Huang, T., Zhang, Y., You, S., Wang, F., Qian, C., Cao, J., and Xu, C. Masked distillation with receptive tokens. *arXiv preprint arXiv:2205.14589*, 2022b.

Huang, Z. and Wang, N. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Kim, B. J., Choi, H., Jang, H., Lee, D. G., Jeong, W., and Kim, S. W. Dead pixel test using effective receptive field. *arXiv preprint arXiv:2108.13576*, 2021.

Kim, J., Park, S., and Kwak, N. Paraphrasing complex network: Network compression via factor transfer. *Advances in neural information processing systems*, 31, 2018.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks.

Advances in neural information processing systems, 25, 2012.

Li, E., Zeng, L., Zhou, Z., and Chen, X. Edge ai: On-demand accelerating deep neural network inference via edge computing. *IEEE Transactions on Wireless Communications*, 19(1):447–457, 2019.

Li, T., Shetty, S., Kamath, A., Jaiswal, A., Jiang, X., Ding, Y., and Kim, Y. Cancergpt: Few-shot drug pair synergy prediction using large pre-trained language models. *arXiv preprint arXiv:2304.10946*, 2023.

Li, Y., Chen, X., Dong, M., Tang, Y., Wang, Y., and Xu, C. Spatial-channel token distillation for vision mlps. In *International Conference on Machine Learning*, pp. 12685–12695. PMLR, 2022.

Liu, S., Chen, T., Chen, X., Chen, X., Xiao, Q., Wu, B., Pechenizkiy, M., Mocanu, D., and Wang, Z. More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity. *arXiv preprint arXiv:2207.03620*, 2022a.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., and Guo, B. Swin transformer v2: Scaling up capacity and resolution. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022b.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. A convnet for the 2020s. *arXiv preprint arXiv:2201.03545*, 2022c.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Luo, W., Li, Y., Urtasun, R., and Zemel, R. Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems*, 29, 2016.

Mao, X., Qi, G., Chen, Y., Li, X., Duan, R., Ye, S., He, Y., and Xue, H. Towards robust vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12042–12051, 2022.

Mirzadeh, S. I., Farajtabar, M., Li, A., Levine, N., Matsukawa, A., and Ghasemzadeh, H. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 5191–5198, 2020.

Park, W., Kim, D., Lu, Y., and Cho, M. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019.

Peng, C., Zhang, X., Yu, G., Luo, G., and Sun, J. Large kernel matters–improve semantic segmentation by global convolutional network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4353–4361, 2017.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., and Bengio, Y. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252, 2015.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

Touvron, H., Bojanowski, P., Caron, M., Cord, M., El-Nouby, A., Grave, E., Izacard, G., Joulin, A., Synnaeve, G., Verbeek, J., et al. Resmlp: Feedforward networks for image classification with data-efficient training. *arXiv preprint arXiv:2105.03404*, 2021a.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357. PMLR, 2021b.

Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., and Jégou, H. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 32–42, 2021c.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, H., Ge, S., Lipton, Z., and Xing, E. P. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019.

Wang, Z., Bai, Y., Zhou, Y., and Xie, C. Can cnns be more robust than transformers? *arXiv preprint arXiv:2206.03452*, 2022.

Wei, Y., Hu, H., Xie, Z., Zhang, Z., Cao, Y., Bao, J., Chen, D., and Guo, B. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *arXiv preprint arXiv:2205.14141*, 2022.

Wightman, R. GitHub repository: Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Wightman, R., Touvron, H., and Jégou, H. Resnet strikes back: An improved training procedure in timm. *arXiv preprint arXiv:2110.00476*, 2021.

Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I. S., and Xie, S. Convnext v2: Co-designing and scaling convnets with masked autoencoders. *arXiv preprint arXiv:2301.00808*, 2023.

Xiao, Q., Wu, B., Zhang, Y., Liu, S., Pechenizkiy, M., Mocanu, E., and Mocanu, D. C. Dynamic sparse network for time series classification: Learning what to "see". In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=ZxOO5jfqSYw.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

Xue, M., Song, J., Wang, X., Chen, Y., Wang, X., and Song, M. Kdexplainer: A task-oriented attention model for explaining knowledge distillation. *arXiv preprint arXiv:2105.04181*, 2021.

Yang, C., Xie, L., Su, C., and Yuille, A. L. Snapshot distillation: Teacher-student optimization in one generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2859–2868, 2019.

Yang, J., Martinez, B., Bulat, A., and Tzimiropoulos, G. Knowledge distillation via softmax regression representation learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=ZzwDy_wiWv.

Yang, T., Zhang, H., Hu, W., Chen, C., and Wang, X. Fastparc: Position aware global kernel for convnets and vits. *arXiv preprint arXiv:2210.04020*, 2022a.

Yang, Z., Li, Z., Gong, Y., Zhang, T., Lao, S., Yuan, C., and Li, Y. Rethinking knowledge distillation via cross-entropy. *arXiv preprint arXiv:2208.10139*, 2022b.

Yang, Z., Li, Z., Jiang, X., Gong, Y., Yuan, Z., Zhao, D., and Yuan, C. Focal and global knowledge distillation for detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4643–4652, 2022c.

Yang, Z., Li, Z., Shao, M., Shi, D., Yuan, Z., and Yuan, C. Masked generative distillation. *arXiv preprint arXiv:2205.01529*, 2022d.

Yang, Z., Li, Z., Zeng, A., Li, Z., Yuan, C., and Li, Y. Vitkd: Practical guidelines for vit feature knowledge distillation. *arXiv preprint arXiv:2209.02432*, 2022e.

Yao, X., ZHANG, Y., Chen, Z., Jia, J., and Yu, B. Distill vision transformers to CNNs via low-rank representation approximation, 2023. URL https://openreview.net/forum?id=U4llPAUi4z.

Yim, J., Joo, D., Bae, J., and Kim, J. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4133–4141, 2017.

Yuan, L., Tay, F. E., Li, G., Wang, T., and Feng, J. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3903–3911, 2020.

Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.-H., Tay, F. E., Feng, J., and Yan, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 558–567, 2021.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019.

Zagoruyko, S. and Komodakis, N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Zhang, J. and Tao, D. Empowering things with intelligence: a survey of the progress, challenges, and opportunities in artificial intelligence of things. *IEEE Internet of Things Journal*, 8(10):7789–7817, 2020.

Zhao, B., Cui, Q., Song, R., Qiu, Y., and Liang, J. Decoupled knowledge distillation. *arXiv preprint arXiv:2203.08679*, 2022.

Zhong, Z., Zheng, L., Kang, G., Li, S., and Yang, Y. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 13001–13008, 2020.

Zhou, H., Song, L., Chen, J., Zhou, Y., Wang, G., Yuan, J., and Zhang, Q. Rethinking soft labels for knowledge distillation: A bias–variance tradeoff perspective. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=gIHd-5X324.

A. FLOPS and Params of Teacher Models and Student Models

We carefully chose the models so that they have roughly a similar model size and FLOPs, i.e., about 5.0G FLOPs and 30M parameter count, as shown in Table 8.

Table 8. **Params and FLOPS of Teacher models and Student models.**

Teacher	Params(M)	FLOPS	Student	Params(M)	FLOPS
ViT-S	22	4.6G	ResNet-50	23	4G
Swin-T	28	4.5G	ResNet-50	23	4G
CSWin-T	23	4.3G	ResNet-50	23	4G
ConvNeXt-T	29	4.5G	ResNet-50	23	4G
SLaK-T	30	5.0G	ResNet-50	23	4G
Swin-T	28	4.5G	ConvNeXt-T	29	4.5G
CSWin-T	23	4.3G	ConvNeXt-T	29	4.5G
ConvNeXt-T	29	4.5G	ConvNeXt-T	29	4.5G
SLaK-T	30	5.0G	ConvNeXt-T	29	4.5G

B. Robustness Evaluation of Distilled ConvNeXt and State-of-the-Art Baselines

We could see from Table 9 that the distilled ConvNeXt student model surpasses the performance of most baseline models and attains a level of robustness comparable to the large-kernel-based model such as SLaK-Tiny, as well as high-capacity models like Swin-B and RVT-B. This strongly suggests that distillation from large-kernel models not only enhances performance but also serves as a robustness booster.

Table 9. **Robustness Evaluation of ConvNeXt and Baselines**. We do not make use of any specialized modules or additional fine-tuning procedures.

Model	Data/Size	#Training Epoches	FLOPs / Params	Clean	C (↓)	A(↑)	R(↑)	SK(↑)
RVT-S* (Mao et al., 2022)	$1K/224^2$	300	4.7 / 23.3	81.9	49.4	25.7	47.7	34.7
ConvNeXt-T (Liu et al., 2022c)	$1K/224^2$	300	4.5 / 28.6	82.1	53.2	24.2	47.2	33.8
ConViT-S (d'Ascoli et al., 2021)	$1K/224^2$	300	5.4 / 27.8	81.5	49.8	24.5	45.4	33.1
SLak-T (Liu et al., 2022a)	$1K/224^2$	300	5.0 / 29	82.5	41.2	29.9	45.3	32.4
Swin-B (Liu et al., 2021)	$1K/224^2$	300	15.4 / 87.8	**83.4**	54.4	**35.8**	46.6	32.4
RVT-B* (Mao et al., 2022)	$1K/224^2$	300	17.7 / 91.8	82.6	46.8	28.5	48.7	36.0
ConVNeXt L2S-T	$1K/224^2$	120	4.5/29	82.17	42.02	24.37	47.50	35.13
ConVNeXt L2S-T	$1K/224^2$	300	4.5/29	83.10	**40.30**	28.46	**49.05**	**36.64**

C. Experiments of ResNet-50 Student with Strong Training Recipe from Timm

To evaluate if large-kernel ConvNets teachers are still better than other architectures under optimal settings, we adopt the optimal training recipe of Timm and report the distilled ResNet-50 in the following table. This training recipe gives us 78.06% top-1 accuracy with ResNet-50 on ImageNet trained for 120 epochs, which matches the Timm A1 one reported in Wightman et al. (2021). i.e., 78.1%.

Results are reported in Table 10. Again, we see that large-kernel teachers, i.e., ConvNeXt-T and SLaK-T, consistently outperform transformer-based teachers in distilling the ResNet-50 under this optimal training recipe. We believe these results support our claim that large-kernel ConvNets are better teachers than Vision Transformers for small-kernel ConvNets.

Table 10. **NKD Results of ResNet-50 Distilled from Various Teachers on ImageNet.** All models are distilled for 120 epochs and are trained with the strong training recipe from Timm.

Teacher	Arch. Type	Student	Teacher Top-1	Student Top-1	Distilled Top-1	Effective Gain	Direct Gain
ViT-S	Transformer	ResNet-50	79.8	78.06	78.50	0.5	0.44
Swin-T	Transformer	ResNet-50	81.3	78.06	78.76	0.8	0.70
CSWin-T	Transformer	ResNet-50	82.7	78.06	78.80	0.89	0.74
ConvNeXt-T	ConvNet	ResNet-50	82.1	78.06	78.95	1.08	0.89
SLaK-T	ConvNet	ResNet-50	82.5	78.06	**79.14**	**1.3**	**1.1**

D. Experiments of MobileNet Student Model

we conducted experiments on MobileNet-V3 and reported the results in Table 11. The results demonstrate that large-kernel models like ConvNeXt-T and SLaK-T consistently surpass all transformer-based teachers in terms of both Effective Gain and Direct Gain when paired with the lightweight model MobileNet. Our results demonstrate that the benefits of large-kernel ConvNets over Vision Transformers can be generalized to lightweight ConvNets.

Table 11. **NKD results of MobileNet distilled from various teachers on ImageNet.** All models are distilled for 120 epochs and are trained with Timm's training recipe.

Teacher	Arch. Type	Student	Teacher Top-1	Student Top-1	Distilled Top-1	Effective Gain	Direct Gain
ViT-S	Transformer	MobileNetV3	79.8	73.59	74.42	1.04	0.83
Swin-T	Transformer	MobileNetV3	81.3	73.59	74.53	1.15	0.94
CSWin-T	Transformer	MobileNetV3	82.7	73.59	74.51	1.11	0.92
ConvNeXt-T	ConvNet	MobileNetV3	82.1	73.59	74.67	1.31	1.08
SLaK-T	ConvNet	MobileNetV3	82.5	73.59	**75.01**	**1.72**	**1.42**

E. Experiments on Pre-trained Model

We investigate the impact of knowledge distillation on the pre-trained step of a student model. We pre-train a student model on ImageNet-1K and fine-tune it on CIFAR-10 and CIFAR-100. We apply knowledge distillation to the ImageNet-1K pre-training step. Results are reported in Table 12 and Table 13.

The Table 12 and Table 13 reveal that pre-trained student models distilled from large-kernel teachers, such as SLaK-T and ConvNeXt-T, achieve higher accuracies in downstream tasks like CIFAR-10/100 compared to those distilled from transformer-based teachers. Notably, the ResNet-50 distilled (by NKD) from SLaK-T outperforms the CSwin-T's counterpart by a significant margin (1.5%) on CIFAR-100, highlighting its benefits. These results demonstrate the effectiveness of knowledge distillation in the pre-training stage, and moreover, large-kernel teachers provide more benefits than Vision Transformers.

Table 12. **The results of fine-tuning the NKD distilled ResNet-50 and ConvNeXt-T models on downstream tasks CIFAR-10/100.** All models are fine-tuned based on SGD optimizer with lr=1e-3 for 50 epochs.

Teacher	Arch. Type	Student	Distilled Top-1 on ImageNet-1K	CIFAR-10	CIFAR-100
Swin-T	Transformer	ResNet-50	77.67	96.1	81.6
CSWin-T	Transformer	ResNet-50	77.68	96.2	82.0
ConvNeXt-T	ConvNet	ResNet-50	78.04	96.6	82.1
SLaK-T	ConvNet	ResNet-50	**78.57**	**97.0**	**83.5**
Swin-T	Transformer	ConvNeXt-T	81.10	97.4	86.9
CSWin-T	Transformer	ConvNeXt-T	81.65	97.7	86.9
ConvNeXt-T	ConvNet	ConvNeXt-T	81.77	97.9	87.2
SLaK-T	ConvNet	ConvNeXt-T	**82.17**	**98.3**	**87.5**

Table 13. **The results of fine-tuning the KD distilled ResNet-50 and ConvNeXt-T models on downstream tasks CIFAR-10/100.** All models are fine-tuned based on SGD optimizer with lr=1e-3 for 50 epochs.

Teacher	Arch. Type	Student	Distilled Top-1 on ImageNet-1K	CIFAR-10	CIFAR-100
Swin-T	Transformer	ResNet-50	76.87	95.88	81.37
CSWin-T	Transformer	ResNet-50	76.77	95.93	81.78
ConvNeXt-T	ConvNet	ResNet-50	76.93	96.15	82.16
SLaK-T	ConvNet	ResNet-50	**77.05**	**96.34**	**82.53**
Swin-T	Transformer	ConvNeXt-T	80.93	97.17	86.22
CSWin-T	Transformer	ConvNeXt-T	81.18	97.41	86.52
ConvNeXt-T	ConvNet	ConvNeXt-T	81.63	97.69	86.96
SLaK-T	ConvNet	ConvNeXt-T	**81.86**	**98.02**	**87.13**