Exploring the Knowledge Transferred by Response-Based Teacher-Student Distillation

Liangchen Song
University at Buffalo
Buffalo, USA

Xuan Gong
University at Buffalo
Buffalo, USA

Helong Zhou
Horizon Robotics
Beijing, China

Jiajie Chen
Horizon Robotics
Beijing, China

Qian Zhang
Horizon Robotics
Beijing, China

David Doermann
University at Buffalo
Buffalo, USA

Junsong Yuan
University at Buffalo
Buffalo, USA

ABSTRACT

Response-based Knowledge Distillation refers to the technique of supervising the student network with the teacher networks' predictions. The method is motivated by observing that the predicted probabilities reflect the relation among labels, which is the knowledge to be transferred. This paper explores the transferred knowledge from a novel perspective: comparing the knowledge transferred through different teachers. Two intriguing properties are observed. First, higher confidence scores of teachers' predictions lead to better distillation results, and second, teachers' incorrectly predicted training samples should be kept for distillation. We then analyze the phenomenon by studying teachers' decision boundaries, of which some can help the student generalize while some may not. Based on the observations, we further propose an embarrassingly simple distillation framework named Efficient Distillation, which is effective on ImageNet with different teacher-student pairs: When using ResNet34 as the teacher, the student ResNet18 trained **from scratch** reaches **74.07%** Top-1 accuracy **within 98 GPU hours** (RTX 3090), outperforming current state-of-the-art result (73.19%) by a large margin. Our code is available at https://github.com/lsongx/EffDstl.

CCS CONCEPTS

• **Computing methodologies** → *Computer vision*; **Neural networks**;

KEYWORDS

Knowledge Distillation, Teacher-Student Learning

ACM Reference Format:
Liangchen Song, Xuan Gong, Helong Zhou, Jiajie Chen, Qian Zhang, David Doermann, and Junsong Yuan. 2023. Exploring the Knowledge Transferred by Response-Based Teacher-Student Distillation. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3612162

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada.

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612162

1 INTRODUCTION

Response-based Knowledge Distillation (KD), introduced in [14], transfers the knowledge contained by a well-trained teacher's predictions network to a student network. KD has been successfully applied to various visual tasks, such as semantic segmentation [20], object detection [2], and human pose estimation [25].

KD is originally built on the insight that the teacher's predicted probabilities encode the correlation among classes. Adopting the predictions as soft targets for supervision helps the student generalize. In this paper, we are concerned with understanding the knowledge contained in the soft targets. Our investigation is inspired by a phenomenon reported by a state-of-the-art (SoTA) distillation framework MEAL v2 [36]: switching the teacher to a model with lower accuracy leads to a significant distillation performance drop. As demonstrated in Fig. 1, the distillation performance gap indicates the existence of different knowledge transferred between the two teachers.

The difference between two kinds of teachers is studied: high performance with slow inference speed teacher, named as a strong teacher, and relatively worse performance with faster inference speed teacher, called a weak teacher. To explore the knowledge transferred by different teachers, we propose to replace the soft targets from the weak teacher with that from the strong teacher with other schemes during training. The replacement is designed in two aspects: replacement for sample subsets and replacement for soft targets.

For sample subsets, two kinds of samples are studied, inspired by previous studies [51, 53]: regularization samples and correct samples, where regularization means that the teacher's confidence score is higher than the student's score, and correct means the teacher makes an accurate prediction. Furthermore, three replacement schemes are designed to progressively explore the knowledge: *full replacement* that changes all values, *partial replacement* that changes relative confidence of non-ground-truth classes, and *confidence replacement* that changes the confidence of the ground-truth class. In our investigation, the soft targets from a weak teacher are

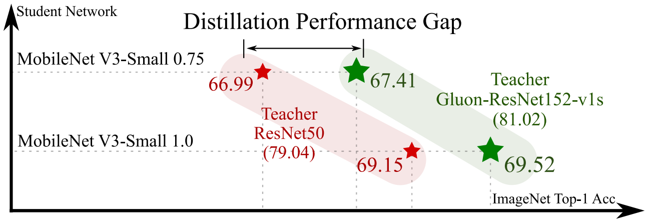


Figure 1: Distillation results (Top-1 accuracy on ImageNet) with different teachers for MEAL V2 [36] framework. The performance of the teacher has a significant impact on the distillation result.

replaced by those from a strong teacher and manually designed labels on different training subsets.

Our exploration reveals two intriguing properties about the distillation scheme proposed in MEAL V2: 1) Replacing the confidence score of the labeled class can boost the distillation performance. Since the confidence of predictions from a strong teacher is higher than that of a weak teacher, we observe that higher confidence scores lead to better distillation results. 2) Replacing soft targets from the teacher by manually designed labels on regularization and not correct samples leads to a significant performance drop. That is, teachers' incorrectly predicted training samples should be kept for distillation.

We hypothesize that the student mimics the teacher's decision boundaries as a reason behind the phenomenon. Some decision boundaries help the student generalize, denoted as *good boundaries* (*good-bd*), while some boundaries are the bias from the teacher [22, 53], denoted as *bad boundaries* (*bad-bd*). Then, samples near the *good-bd* should be kept while samples near the *bad-bd* can be manipulated. We demonstrate networks' predictions on adversarial examples near the two boundaries as proof-of-concept experiments.

Finally, based on the observations, we propose a distillation framework named efficient distillation (EffDstl). The critical component in the framework is that we enforce the teacher's confidence score consistently higher than the student's score for the *bad-bd* samples while keeping distillation on the *good-bd* samples unchanged. Since the subset of samples is dynamically determined during training, we name the simple technique consistent knowledge mining. With consistent knowledge mining, a weak teacher can be converted to a strong teacher, reducing the distillation cost of querying a teacher's soft targets. A straightforward comparison between our EffDstl and MEAL V2 is demonstrated in Fig. 2. With the proposed EffDstl and ResNet34 as the teacher, the student network ResNet18 can be trained from scratch and reaches 74.07% on ImageNet [3], while remarkably reducing distillation costs. To sum up, our contributions are as follows:

- We explore the knowledge transferred on the large-scale dataset ImageNet by comparing the distillation performance of a strong teacher and a weak teacher. Two intriguing properties are found with SoTA distillation method MEAL V2 [36]: 1) higher confidence in predictions leads to better distillation for some samples, 2) the student should follow the teacher if the teacher's prediction is different from the given label.

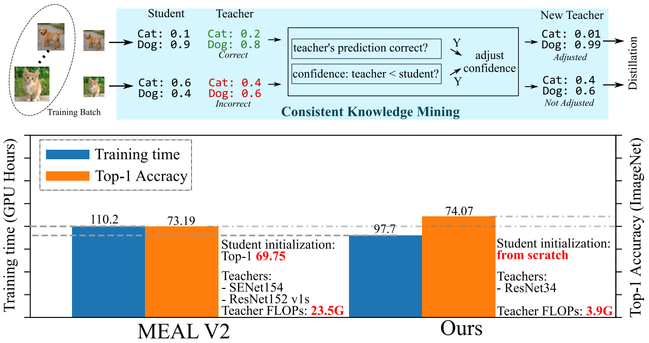


Figure 2: The comparison of distillation results between our Efficient Distillation (EffDstl) and the current SoTA MEAL V2. ResNet18 is chosen as the student.

- We analyze the observed phenomenon from the perspective of decision boundaries: samples that should be distilled are near the teacher's decision boundaries that can help the student generalize, while samples far away from such boundaries can be manipulated. Our results provide a new perspective for understanding knowledge distillation.
- Motivated by our observations, a distillation framework named efficient distillation is developed, and consistent knowledge mining is proposed to convert a weak teacher to a strong teacher. Our framework reaches a new SoTA distillation performance. With our framework, classical manually designed networks like MobileNet V2 outperform advanced lightweight networks, including transformer-based and neural-architecture-search-based networks.

2 RELATED WORK

Knowledge distillation. Knowledge distillation has been widely studied in recent years. In a recent review [10], the knowledge used for distillation is categorized into three groups: response-based knowledge, Feature-based knowledge, and relation-based knowledge. Our method can be classified into response-based knowledge distillation. Response means the final output class-wise probability score of the network. Response-based knowledge is first introduced by [14], in which the distillation loss is defined as the Kullback-Leibler divergence between the student's prediction and the teacher's prediction. The intermediate feature maps are used to supervise the student for feature-based knowledge and are first introduced in FitNets [32]. Feature-based knowledge then further improved by a lot of works, such as attention in [50]. , SVD in [17], AB in [13], ALP in [28] and SemCKD in [1]. The Gram matrix of feature maps from different layers is used as the knowledge in [47]. In [27], mutual information flow from pairs of feature maps is defined as knowledge. Relations contained in samples are also widely used, such as RKD [26], IRG [19], and SP [43].

Investigation of soft targets based knowledge. Besides training with the soft targets generated from a teacher network, several methods are proposed to directly generate the soft targets, such as label smoothing [29, 38]. The connection between soft targets designed manually and generated by the teacher received a lot of attention recently. In [24], the authors demonstrate that teachers trained with label smoothing techniques are not suitable for distillation. However, more recently, in [35], empirical results show that label smoothing does not always suppress the effectiveness of KD. The connection is also studied in [49], and the authors found the regularization property of the soft targets. Analyses of sample-wise soft label-based knowledge are presented by [34, 39, 40, 45, 51]. Our work is more related to [53], in which the authors propose to reduce the weight of the samples that the teacher performs worse than the student. There are mainly three differences between our paper and WSL [53]: 1) We analyze the knowledge by comparing a weak and a strong teacher, while WSL analyzes the knowledge of the same teacher on different samples; 2) We find that the incorrectly predicted samples should be kept unchanged for distillation, which is not observed before; 3) We propose a method to enhance the knowledge of a weak teacher.

Understanding teacher-student distillation. Our exploration reveals intriguing properties about teacher-student distillation since some training samples can be ignored while achieving better distillation performance. Lopez-Paz *et al.* [21] proves that distillation is helpful when soft targets can speed up training. In [30], the effectiveness of distillation is proven by analyzing the optimization process. Mobahi *et al.* [23] and Zhang *et al.* [52] analyzed the reason behind the effectiveness of self-distillation. Hsu *et al.* [16] investigated the generalization bounds of the student network. In [22], a bias-variance perspective on distillation is provided, and a formal criterion as to what constitutes a "good" teacher is provided. Our work does not provide a theoretical explanation of the effectiveness of distillation, but we provide novel intuition about the knowledge transferred by response-based teacher-student distillation.

3 PRELIMINARIES

For KD, we aim to train a student network s under the guidance of a teacher network t. For each sample x associated with a label y, we denote the ith logits output of network s as z_i^s, then the predicted probability for kth class of s is $\hat{y}_{k,\tau}^s = \frac{e^{z_k^s/\tau}}{\sum_i e^{z_i^s/\tau}}$, where τ is a scaling parameter named temperature. Similarly, $\hat{y}_{k,\tau}^t$ denotes the predicted probability for kth class of t. Then the KD loss of this sample is defined as $L_{\text{KD}} = -\tau^2 \sum_k \hat{y}_{k,\tau}^t \log \hat{y}_{k,\tau}^s$. Moreover, cross-entropy loss, $L_{\text{CE}} = -\sum_k y \log \hat{y}_{k,1}^s$, is commonly used for supervising the student and the total loss is a balanced sum of the two losses $L = (1-\lambda)L_{\text{CE}} + \lambda L_{\text{KD}}$, where λ is a balancing weight. In [14], λ is manually tuned to be 0.9. The practice of manual tuning is followed by works like [41]. In [51, 53], the authors propose dynamically determining λ for each training sample. Moreover, MEAL V2 [36] sets $\lambda = 1$ and discards the cross-entropy loss during distillation.

4 KNOWLEDGE FROM DIFFERENT TEACHERS

In Fig. 1, we can observe that switching to a weak teacher (ResNet50 with Top-1 79.04%) leads to a drop of 0.4% distillation accuracy. To figure out why such a performance gap exists, we explore the knowledge transferred by a strong teacher and a weak teacher. Inspired by previous analyses of distillation [24, 35, 53], the exploration is conducted to answer the following two questions: 1) Which value in the teacher output $[\hat{y}_{1,\tau}^t, ..., \hat{y}_{n,\tau}^t]$ leads to the distillation gap? 2) Which sample during training leads to the distillation gap?

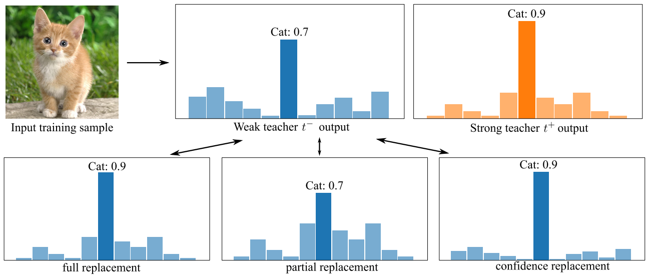
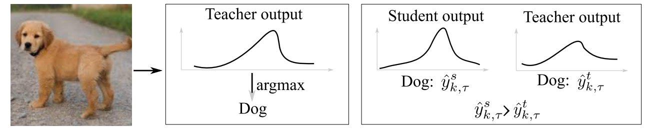


Figure 3: To investigate the difference in knowledge between a weak and a strong teacher, we design three schemes for adjusting the knowledge of the weak teacher based on the strong teacher.


Teacher output
argmax
Dog
Student output
Teacher output
Dog: $\hat{y}_{k,\tau}^s$
Dog: $\hat{y}_{k,\tau}^t$
$\hat{y}_{k,\tau}^s > \hat{y}_{k,\tau}^t$

Input training sample (a) correct sample (b) regularization sample

Figure 4: We study how the two sets of samples affect the distillation performance. A correct sample means the teacher makes an accurate prediction, and a regularization sample means the student generates higher confidence than the teacher.

4.1 Replacing the soft targets

To answer the two questions, our principle is to adjust the predictions of a weak teacher t^- according to a strong teacher t^+ based on different rules. For the first question, we design three alteration schemes: full replacement, partial replacement, and confidence replacement. An illustration of the three alteration schemes of t^- is demonstrated in Fig. 3. Mathematically, for an input labelled as kth class, the distillation label becomes $[\hat{y}_{1,\tau}^{t^+}, ..., \hat{y}_{n,\tau}^{t^+}]$ for full replacement, which means simply using outputs from t^+ as the label. For partial replacement, the ith ($i \neq k$) label becomes

$$\hat{y}_{i,\tau}^t \leftarrow \hat{y}_{i,\tau}^{t^+} \cdot \frac{1 - \hat{y}_{k,\tau}^{t^-}}{1 - \hat{y}_{k,\tau}^{t^+}}. \tag{1}$$

Partial replacement means that we keep the confidence score of the ground-truth label and use the distribution of other classes from t^+ as the supervision. For confidence replacement, the ith ($i \neq k$) label becomes

$$\hat{y}_{i,\tau}^t \leftarrow \hat{y}_{i,\tau}^{t^-} \cdot \frac{1 - \hat{y}_{k,\tau}^{t^+}}{1 - \hat{y}_{k,\tau}^{t^-}}. \tag{2}$$

Unlike partial replacement, confidence replacement keeps the distribution on other classes and uses the confidence score on kth class from y^+.

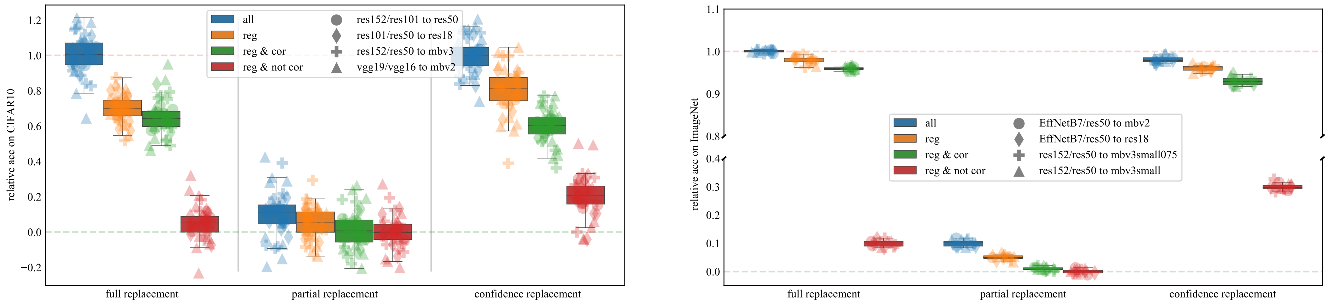


Figure 5: Exploration of replacing the soft labels from a weak teacher to a strong teacher. Three replacement schemes are tested (*c.f.*, Fig. 3). For each weak/strong teacher and student triplet, denote the result with the weak teacher only and strong teacher only as acc^- and acc^+. Then a distillation result acc is normalized by $\frac{\mathrm{acc}-\mathrm{acc}^-}{\mathrm{acc}^+-\mathrm{acc}^-}$, so that we can explore different teacher-student settings in the same figure. We repeat distillation experiments several times and each point in the figure denotes one distillation result. Results on CIFAR10 and ImageNet are demonstrated. The colors denote the groups of samples to that replacements are applied. The results demonstrate that *the confidence of predictions* (*c.f.*, Fig. 3) is more important than class-wise relations for being a good teacher.

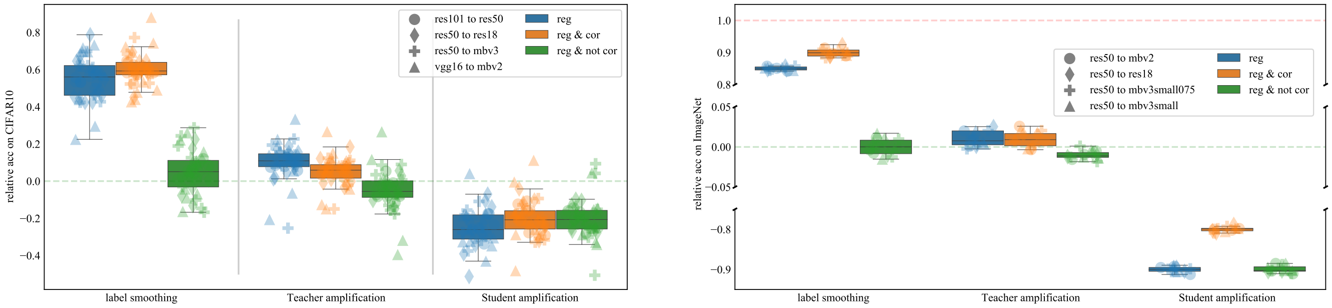


Figure 6: Exploration of manually manipulating the soft labels from a weak teacher. Three manipulation schemes from Eqs. (3) to (5) are investigated. Similar to Fig. 5, distillation is repeated with different pairs and then normalized. The results demonstrate that augmenting the weak teacher with the *label smoothing* scheme on the *reg & cor samples* (*c.f.*, Fig. 4) can approximate the performance of a strong teacher.

For the second question, distillation on different sample subsets is conducted. Two kinds of samples, inspired by the empirical results in [51, 53], are of interest: correct samples and regularization samples. As illustrated in Fig. 4, a sample labeled as kth class is called a correct sample if $\hat{y}^t_{k,\tau} > \hat{y}^t_{i,\tau} (\forall i \neq k)$. Regularization sample follows the idea in [53] and is defined by $\hat{y}^s_{k,\tau} > \hat{y}^t_{k,\tau}$.

Fig. 5 demonstrates the distillation results with different alteration schemes and specific sample subsets combinations. The "not correct" sample denotes the sample on which the teacher makes a wrong prediction. As a response to the two questions asked above, two intriguing properties can be observed from the table: First, full replacement and confidence replacement act similarly; Second, only replacing samples from the "regularization&correct" subset leads to good results. The first property indicates that the confidence of prediction matters most, then a question naturally arises: Can we increase the confidence of prediction and convert a weak teacher into a strong teacher? In the next section, we study several strategies for manipulating soft targets.

4.2 Manipulating the soft targets

Since previous experiments in Fig. 5 suggest that the confidence score matters, a straightforward idea is to increase the confidence scores of teacher outputs. Similar to the experiments conducted in the last section, we design three manipulation baselines: label smoothing, teacher amplification, and student amplification. Label smoothing, proposed in [38], changes the label of a sample labeled as kth class to

$$\hat{y}^t_k \leftarrow t, \quad \hat{y}^t_i \leftarrow \frac{(1-t)}{N-1}, \tag{3}$$

where N is the number of classes and t is a predefined number. Teacher amplification increases the confidence of kth class by a certain value $\delta > 0$, that is,

$$\hat{y}^t_k \leftarrow \max\{\hat{y}^t_k + \delta, 1\}, \quad \hat{y}^t_i \leftarrow \frac{\hat{y}^t_i(1-\hat{y}^t_k)}{1-\hat{y}^t_k-\delta}. \tag{4}$$

Student amplification is similar to teacher amplification, except that the output of students is amplified,

$$\hat{y}^t_k \leftarrow \max\{\hat{y}^s_k + \delta, 1\}, \quad \hat{y}^t_i \leftarrow \frac{\hat{y}^s_i(1-\hat{y}^s_k)}{1-\hat{y}^s_k-\delta}. \tag{5}$$

Results are demonstrated in Fig. 6, from which we find that label smoothing outperforms the other two simple manipulations. Another noteworthy point is that the label smoothing manipulation on the "regularization&correct" subset achieves the best result, approximating the strong teacher-based distillation. In comparison, teacher amplification and student amplification cannot enhance the knowledge or even lead to worse results.

After being manipulated with label smoothing, the regularization samples are not regularized anymore, that is $\hat{y}^s_k < t$. In other words, by applying label smoothing to the "regularization&correct" subset, we enforce that the correct predictions from the teachers output higher confidence than the student.

4.3 Understanding difference of knowledge

Existing theories [16, 22, 30] have proved that the student mimics the teacher, and the student converges to favorable optima after distillation. We hypothesize that the student mimics the teacher's decision boundaries during training. The teacher's decision boundaries are approximated by mimicking the teacher's response to different training samples. We further hypothesize that some decision boundaries of the teacher can help the student generalize, and these boundaries are denoted as good boundaries. The other boundaries can not help the student generalize, and we denote them as bad boundaries. Then during training, a sample should be distilled if it is close to the *good-bd*, and a sample can be manipulated if it is close to the *bad-bd*.

Theoretical clues. Existing theories in [22] shed light on how to determine if a teacher is good for distillation. For a predictor s, we denote its risk (*i.e.*, generalization error) as $R(s)$ and its empirical risk (*i.e.*, training error) on dataset D as $\hat{R}(s; D)$, then with symbols defined in Sec. 3 and for any bounded loss ℓ, we have the following result (Proposition 3 in [22]),

$$\begin{aligned}\mathbb{E}[(R(s)-\hat{R}(s;D))^2] \leqslant \\ \frac{1}{|D|}\mathbb{V}[t(x)^T\ell(s(x))] + O\left(\mathbb{E}[\|t-t^*\|_2]\right)^2,\end{aligned} \tag{6}$$

where $\mathbb{V}$ is the variance and t^* is the *Bayes class-probability distribution*. In Eq. (6), the left side ($\mathbb{E}[(R(s)-\hat{R}(s;D))^2]$) is the gap between training and testing and can be bounded by the sum of two terms: The first one $\frac{1}{|D|}\mathbb{V}[t(x)^T\ell(s(x))]$ quantifies the difference between student and the teacher. The second term $O\left(\mathbb{E}[\|t-t^*\|_2]\right)^2$ measures the difference between the teacher and the Bayes class-probabilities. Exactly assessing Bayes class-probabilities is infeasible in practice, but we can assume that t^* is generally consistent with the given label (*i.e.*, clean annotations). Therefore, if the teacher's prediction becomes inconsistent with t^*, the teacher is introducing a *bad-bd* to the student. In our exploration, we find that *a correct prediction with low confidence* is close to *bad-bd* and thus making the term $O\left(\mathbb{E}[\|t-t^*\|_2]\right)^2$ larger.

Proof-of-concept experiments. Experiments in Figs. 5 and 6 demonstrate that some samples should be kept and some samples should be manipulated to achieve a better distillation performance. We presume that the kept samples are near the *good-bd* and manipulated samples are near the *bad-bd* or far from boundaries. Adversarial attack method FGSM [9] is employed to study the distance between the samples and the boundaries. We randomly select 1000 samples of the kept and manipulated subset and demonstrate the attack results in Fig. 7. We can first observe that confidence in manipulated samples is less sensitive to adversarial examples. Thus these samples are relatively far from decision boundaries. Also, when applying the attack on other networks (student and an undistilled MobileNet), the impact gap between the two sample subsets (peaks of the red and blue lines) becomes larger. This indicates that other networks share boundaries close to those of the kept samples. Furthermore, in Fig. 8, we demonstrate the impact of not keeping the incorrectly predicted training samples during distillation. We can observe that the gap between training accuracy and validation accuracy increases when incorrectly predicted training samples are not kept, which indicates that the student starts to overfit the training set.

5 EFFICIENT DISTILLATION

Based on the previous exploration, we find that the confidence score of predictions is an important difference between a weak and a strong teacher. Besides, the scores can be manually manipulated to enhance the teacher. In this section, we convert our findings into a detailed distillation framework.

5.1 Consistent Knowledge Mining

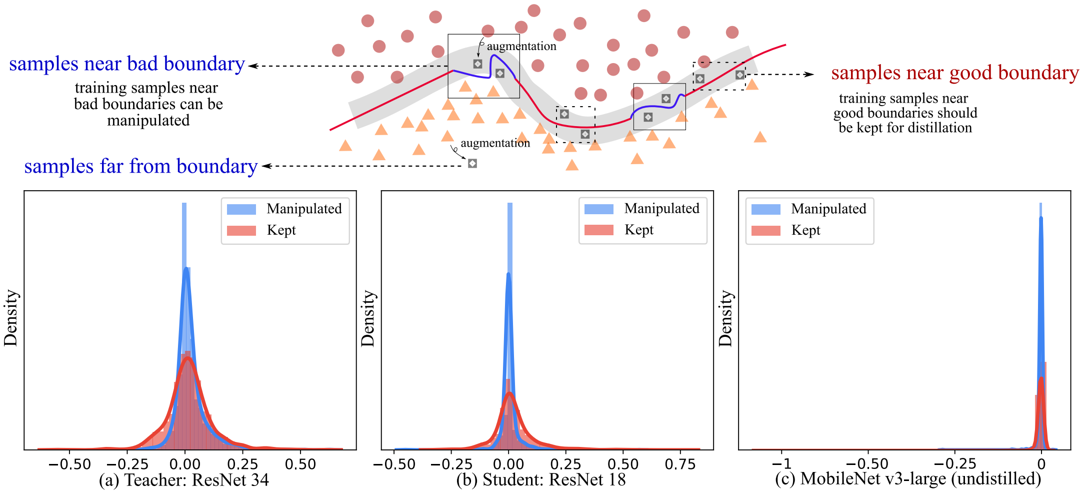
During training, for a sample labeled as kth class, if the teacher make a correct prediction, i.e., $\hat{y}^t_{k,\tau} > \hat{y}^t_{i,\tau}(\forall i \neq k)$, and the prediction confidence is lower than that of the student, i.e., $\hat{y}^t_{k,\tau} < \hat{y}^s_{i,\tau}$, we manually manipulate the distillation label and make the distilled knowledge consistent with provided labels. For manipulation, we use label smoothing with high confidence (0.99) to replace the soft target provided by the teacher. Distillation on other samples remained unchanged. The simple technique is named consistent knowledge mining.

5.2 Framework

We propose testing the distillation performance with improved training hyper-parameters for the classification task, widely adopted for evaluating Neural Architecture Search [6]. Specifically, cosine learning rate decay and more training epochs (300 epochs in total) are adopted in our training. Only "RandomResizedCrop" and "RandomHorizontalFlip" are used for data augmentation, and input resolution is 224×224.

EffDstl consists of a two-stage optimization strategy: The first stage takes 240 epochs with weight decay 4e-5, and the network is trained with WSL distillation loss [53]; The second stage takes 60 epochs with weight decay 0, and the network is trained with our proposed consistent knowledge mining scheme.

Some other hyper-parameters are also tuned for the two stages separately. For the first stage, the initial learning rate is 0.5, and the batch size is 1024. For the second stage, the learning rate is 5e-2



(a) Teacher: ResNet 34 (b) Student: ResNet 18 (c) MobileNet v3-large (undistilled)

Figure 7: Change of prediction confidence under adversarial attack for different distillation subsets. Adversarial examples are generated from the teacher network.

Teacher: ResNet34 (TV, 73.31), ResNet34d (Timm, 77.12)
Student: ResNet18

Method	Top-1 Acc		Top-5 Acc	
	Advanced	(100-epochs)	Advanced	(100-epochs)
No Distill	72.30	(69.75)	90.84	(89.07)
KD [14]	72.90	(70.67)	91.18	(90.04)
RKD [26]	72.19	(70.40)	90.78	(89.78)
WSL [53]	73.44	(72.04)	91.22	(90.70)
EffDstl	**74.07**	(72.17)	**91.54**	(90.95)

Teacher: ResNet50 (TV, 76.16), ResNet50 (Timm, 79.04)
Student: MobileNet V1

Method	Top-1 Acc		Top-5 Acc	
	Advanced	(100-epochs)	Advanced	(100-epochs)
No Distill	73.29	(68.87)	91.38	(88.76)
KD [14]	74.50	(70.49)	92.08	(89.92)
RKD [26]	74.08	(68.50)	91.74	(88.32)
WSL [53]	74.78	(71.52)	91.89	(90.34)
EffDstl	**75.49**	(71.91)	**92.32**	(90.38)

Table 1: A comparison with other knowledge distillation methods. "Advanced" denotes the advanced training hyper-parameters (300 epochs with cosine learning rate decay). "100-epochs" denotes the commonly used training hyper-parameters for evaluating distillation methods, and the networks are not fully trained. Teachers with "TV" denote the models provided by Torchvision, and teachers with "Timm" denote the models provided by Timm.

at the beginning and decreases to 5e-4 after 60 epochs with cosine decay. The temperature τ for the first stage distillation is set to 2. For the second stage, the temperature for the teacher is set to 0.7, and the temperature for the student is set to 1. Moreover, we change the teacher network to a better network with the same architecture for the second stage. For example, we change the ResNet34 provided by Torchvision to ResNet34d provided by Timm [46].

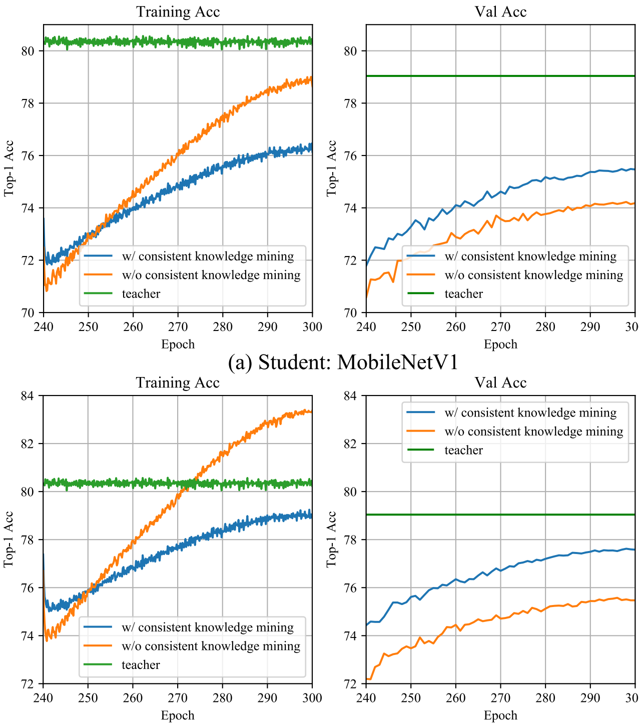
6 EXPERIMENTS

Our experiments are all conducted on ImageNet [4], a large-scale benchmark for classification. There are 1.2 million training images and 50,000 validation images in ImageNet, and the number of classes is 1,000. All training are conducted on 4 NVIDIA RTX 3090 GPUs with batch size 256 on each GPU. For calculating the GPU Hours, we sum the training time of all 4 GPUs.

6.1 Comparison with Other Distillation Methods

We first compare our method with other distillation methods in Tab. 1. Five recently proposed methods are selected for comparison: KD [14], RKD [26], and WSL [53]. In the table, we report the training hyper-parameters with 100 epochs and step learning rate decay to demonstrate the necessity of using the advanced training hyper-parameters. With the advanced training hyper-parameters, ResNet18 and MobileNet V1 trained without distillation outperform all previously reported distillation methods. There is even a large margin between the directly trained MobileNet V1 and the distilled with the 100-epochs hyper-parameters. Such significant gaps make the distillation methods not appealing to practitioners who are looking for compact models.

As reported in Tab. 1, ResNet18 and MobileNet V1 trained with EffDstl are much better than all existing methods. We note that the compared methods are reproduced from the released code from the



(a) Student: MobileNetV1

(b) Student: ResNet34d

Figure 8: Comparison of distillation results with and without keeping the incorrectly predicted samples. Manipulating the incorrectly predicted samples from the teacher leads to a higher training accuracy but lower testing accuracy.

authors with some balancing parameter tuning. Besides, for comparison methods, we also change the teacher network to a better one for the last 60 epochs. We admit that there is inconsistency among the comparisons since we use a two-stage training scheme while the other does not. Thus we do not claim that the previous frameworks, i.e., feature-based and relation-based, are inferior to the proposed EffDstl, but we want to highlight that in-depth analyses are needed for the comparison methods to achieve comparable performance.

6.2 Comparison with SoTA Lightweight Networks

KD is frequently used for acquiring a compact network and is originally proposed for model compression. In Tab. 2, we compare networks distilled with other diligent lightweight network solutions: MobileNet series [15, 33], Tokens-to-token ViT [48], RegNet [31], GhostNet [11], Single-Path NAS [37], DenseNAS [7], Gluon ResNets [12], SK ResNets [18], RepVGG [5], HRNet [44] and Res2Net [8]. The teachers for **EffDstl** - MbNetV3Small0.75 and **EffDstl** - SKResNet18 are ResNet34 (Top-1: 73.31) for the first stage and ResNet34d (Top-1: 77.12) for the second stage. The teachers for **EffDstl** - MobileNet V2, **EffDstl** - ResNet34 and **EffDstl** - ResNet34d are ResNet50 (Top-1: 76.16) for the first stage and ResNet34d (Top-1: 79.04) for the second stage.

Method	Params (M)	FLOPs (M)	Top-1
MobileNet V3-Small 0.75 [15]	2.04	43.4	65.72
MobileNet V3-Small 1.0 [15]	2.54	56.5	67.92
EffDstl - MbNetV3Small0.75	**2.04**	**43.4**	**67.34**
MobileNet V2 [33]	3.50	300.8	72.97
T2T-ViT-7-Distilled [48]	4.30	1300.0	73.10
GhostNet [11]	5.18	141.2	73.98
DeiT-tiny distilled [42]	5.72	1080.1	74.51
DenseNAS-B [7]	4.77	313.6	74.55
EffDstl - MobileNet V2	**3.50**	**300.8**	**74.45**
Gluon-ResNet18-v1b [12]	11.69	1814.1	70.84
ResNet18d [46]	11.71	2053.7	72.26
SKResNet18 [18]	11.96	1814.3	73.04
EffDstl - SKResNet18	**11.96**	**1814.3**	**75.12**
RepVGG-A2 [5]	28.21	5685.3	76.46
HRNet-w18 [44]	21.30	4284.0	76.76
SKResNet34 [18]	22.28	3664.2	76.91
EffDstl - ResNet34	**21.80**	**3663.8**	**77.06**
Res2Net50-48w-2s [8]	25.29	4159.8	77.52
Gluon-ResNet50-v1b [12]	25.56	4089.2	77.58
EffDstl - ResNet34d	**21.82**	**3903.4**	**77.62**

Table 2: Comparison with SoTA lightweight networks.

	Student	Top-1	Training Time
MEAL V2	MobileNet V3-Small 0.75	67.60	103.1 (+?)
	ResNet18	73.19	110.2 (+?)
EffDstl	MobileNet V3-Small 0.75	67.34	93.0
	ResNet18	74.07	97.7
	SKResNet18	75.12	127.3
	MobileNet V1	75.49	111.5
	ResNet34	77.06	160.8

Table 3: Comparison of the distillation costs. Training time is GPU Hours measured with RTX 3090. For MEAL V2, (+?) **indicates the unknown training time for pre-training the student network. As a comparison, we train the student *from scratch* thus, no extra training time is needed.**

The practical value of our proposed method can be observed from Tab. 2. With EffDstl, the networks proposed years ago perform better than some of the most advanced networks. For example, our distilled MobileNet V2 outperforms the transformer-based network T2T-ViT. Also, vanilla ResNet18 and ResNet34 boosted by EffDstl surpasses ResNet34 and ResNet50 from GluonCV [12]. Our proposed distillation is compatible with other lightweight network enhancing techniques, which can be observed from the ResNet34 and ResNet34d pair. Our method provides a new baseline for evaluating lightweight networks.

Teacher		Top-1
Stage 1	Stage 2	
ResNet34 (73.31)	ResNet34d (77.12)	67.34
MobileNet V2 (72.97)	ResNet34d (77.12)	66.63
ResNet34 (73.31)	MobileNet V3-Large (75.52)	67.52
MobileNet V2 (72.97)	MobileNet V3-Large (75.52)	67.18

(a) The impact different teachers. (Student: MbNet V3-Small 0.75)

Student	Top-1		Top-5	
	Before	After	Before	After
MobileNet V2	74.45	74.76	91.74	92.03
ResNet18	74.07	74.46	91.54	91.61
ResNet18d	74.98	75.16	91.97	92.17
SKResNet18	75.12	75.32	92.14	92.24

(b) Further distillation with MEAL V2.

Change Teacher	Student	Top-1
✗	ResNet18	73.36
✓	ResNet18	74.07
✗	MobileNetV1	74.52
✓	MobileNetV1	75.49

(c) Ablation of two stage training.

Stage 2 τ	Student	Top-1
1.0	ResNet18	73.99
0.8	ResNet18	74.24
1.0	MobileNetV1	74.52
0.8	MobileNetV1	75.46

(d) Ablation of τ for stage 2.

Consistent Knowledge Mining	Student	Top-1
✗	MobileNetV1	74.23
✓	MobileNetV1	75.49
✗	ResNet34d	75.58
✓	ResNet34d	77.62

(e) Ablation of consistent knowledge mining.

Table 4: Ablation studies of our proposed Efficient Distillation framework.

6.3 Distillation Costs

One of the main contributions of our method is that we do not rely on large teacher networks, thus significantly reducing the distillation costs while keeping SoTA performance. In Tab. 3, we demonstrate the distilled cost of our method and include the training time of MEAL V2 as a comparison.

The advantage of our EffDstl is obvious. We train the student from scratch, and the final result is comparable to or better than MEAL V2, which needs a well-trained student for initialization. With EffDstl, ResNet18 can be trained from scratch and achieve 74.07 in about 24 hours if 4 RTX 3090 GPUs are available. Also, ResNet34 achieves 77.06 after 1.5 days if trained on 4 RTX 3090 GPUs. These results highlight the efficiency of our proposed EffDstl.

6.4 Ablation Studies

Since different teachers are used for different students in Tab. 2, we first study the impact of using different teachers for distillation. The ablation results are demonstrated in Tab. 4 (a). We find that using a slightly worse teacher MobileNet V2 in stage 1 leads to a much worse result, perhaps due to the label smoothing trick used when training the teacher, so a smaller temperature is needed to make the teacher's prediction sharper. Another possibility is that if the teacher networks of the two stages are of the same type, the student achieves a better result.

Consequently, as our approach draws a significant inspiration from the MEAL V2 technique, we have conducted an investigation to determine whether our method attains a similar objective as MEAL V2. To achieve this, we have subjected the distillation outputs from EffDstl to MEAL V2, as illustrated in Tab. 4 (b). It is crucial to highlight that MEAL V2 employs large teachers, whereas our distillation teachers are either ResNet34 or ResNet50. From the table, it is evident that the Top-1 performance of the students improves by approximately 0.3 for MobileNet V2 and ResNet18. However, the improvements are less significant when compared to the distillation outcomes documented in MEAL V2.

7 DISCUSSION

Limitation. The motivation for this study stems from the utilization of response-based distillation in MEAL V2. Other forms of distillation, such as feature-based and relation-based distillation, have not been thoroughly examined in this research. It is necessary to investigate how the observed properties extend to these alternative distillation frameworks. Additionally, even though we have supplied proof-of-concept experiments and offered intuitive explanations for the observations, our analyses of the phenomenon could be further supported by more in-depth theoretical underpinnings.

Conclusion. In this paper, we investigate the transfer of knowledge via response-based distillation from a weak and a strong teacher in order to further understanding the term "knowledge". We make two notable observations. Firstly, the strong teacher generates greater prediction confidence than the weak teacher, which results in better distillation outcomes. Secondly, low confidence samples that have been accurately predicted by the teachers can be amplified. We provide an insightful analysis of this phenomenon and demonstrate that samples that should be retained are located closer to optimal decision boundaries. Lastly, we introduce a distillation framework, named EffDstl that is remarkably simple and only requires a weak teacher. We show that, using EffDstl, efficient classical network models such as MobileNet V2 can outperform transformer-based or NAS-based networks.

ACKNOWLEDGMENTS

This work is supported in part by a gift grant from Horizon Robotics and the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001120C0124. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA).

REFERENCES

[1] Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen. 2021. Cross-layer distillation with semantic calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 7028–7036.

[2] Guobin Chen, Wongun Choi, Xiang Yu, Tony X. Han, and Manmohan Chandraker. 2017. Learning Efficient Object Detection Models with Knowledge Distillation. In *Advances in Neural Information Processing Systems*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 742–751.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A large-scale hierarchical image database. In *The IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 248–255.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *The IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 248–255.

[5] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. 2021. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13733–13742.

[6] Jiemin Fang, Yuzhu Sun, Qian Zhang, Yuan Li, Wenyu Liu, and Xinggang Wang. 2020. Densely connected search space for more flexible neural architecture search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[7] Jiemin Fang, Yuzhu Sun, Qian Zhang, Yuan Li, Wenyu Liu, and Xinggang Wang. 2020. Densely connected search space for more flexible neural architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10628–10637.

[8] Shanghua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip HS Torr. 2019. Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence* (2019).

[9] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*, Yoshua Bengio and Yann LeCun (Eds.).

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

[11] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. 2020. Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1580–1589.

[12] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. 2019. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 558–567.

[13] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. 2019. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 3779–3787.

[14] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the Knowledge in a Neural Network. *CoRR* abs/1503.02531 (2015).

[15] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. 2019. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1314–1324.

[16] Daniel Hsu, Ziwei Ji, Matus Telgarsky, and Lan Wang. 2021. Generalization bounds via distillation. In *International Conference on Learning Representations*. OpenReview.net.

[17] Seung Hyun Lee, Dae Ha Kim, and Byung Cheol Song. 2018. Self-supervised knowledge distillation using singular value decomposition. In *European Conference on Computer Vision*. 335–350.

[18] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. 2019. Selective kernel networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 510–519.

[19] Yufan Liu, Jiajiong Cao, Bing Li, Chunfeng Yuan, Weiming Hu, Yangxi Li, and Yunqiang Duan. 2019. Knowledge distillation via instance relationship graph. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7096–7104.

[20] Yifan Liu, Changyong Shu, Jingdong Wang, and Chunhua Shen. 2020. Structured Knowledge Distillation for Dense Prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020), 1–1. https://doi.org/10.1109/TPAMI.2020.3001940

[21] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. 2016. Unifying distillation and privileged information. In *International Conference on Learning Representations*, Yoshua Bengio and Yann LeCun (Eds.).

[22] Aditya K Menon, Ankit Singh Rawat, Sashank Reddi, Seungyeon Kim, and Sanjiv Kumar. 2021. A statistical perspective on distillation. In *International Conference on Machine Learning*. PMLR, 7632–7642.

[23] Hossein Mobahi, Mehrdad Farajtabar, and Peter L. Bartlett. 2020. Self-Distillation Amplifies Regularization in Hilbert Space. In *Advances in Neural Information Processing Systems*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.).

[24] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help?. In *Advances in Neural Information Processing Systems*. 4694–4703.

[25] Xuecheng Nie, Yuncheng Li, Linjie Luo, Ning Zhang, and Jiashi Feng. 2019. Dynamic Kernel Distillation for Efficient Pose Estimation in Videos. In *IEEE/CVF International Conference on Computer Vision*. IEEE, 6941–6949.

[26] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. 2019. Relational knowledge distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3967–3976.

[27] Nikolaos Passalis, Maria Tzelepi, and Anastasios Tefas. 2020. Heterogeneous knowledge distillation using information flow modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2339–2348.

[28] Peyman Passban, Yimeng Wu, Mehdi Rezagholizadeh, and Qun Liu. 2021. ALP-KD: Attention-Based Layer Projection for Knowledge Distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 13657–13665.

[29] Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. 2017. Regularizing Neural Networks by Penalizing Confident Output Distributions. In *International Conference on Learning Representations Workshop*. OpenReview.net.

[30] Mary Phuong and Christoph Lampert. 2019. Towards understanding knowledge distillation. In *International Conference on Machine Learning*. 5142–5151.

[31] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. 2020. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10428–10436.

[32] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2015. FitNets: Hints for Thin Deep Nets. In *International Conference on Learning Representations*.

[33] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *The IEEE Conference on Computer Vision and Pattern Recognition*. 4510–4520.

[34] Muhamad Risqi U Saputra, Pedro PB de Gusmao, Yasin Almalioglu, Andrew Markham, and Niki Trigoni. 2019. Distilling knowledge from a deep pose regressor network. In *Proceedings of the IEEE International Conference on Computer Vision*. 263–272.

[35] Zhiqiang Shen, Zechun Liu, Dejia Xu, Zitian Chen, Kwang-Ting Cheng, and Marios Savvides. 2020. Is Label Smoothing Truly Incompatible with Knowledge Distillation: An Empirical Study. In *International Conference on Learning Representations*.

[36] Zhiqiang Shen and Marios Savvides. 2020. MEAL V2: Boosting Vanilla ResNet-50 to 80%+ Top-1 Accuracy on ImageNet without Tricks. *arXiv preprint arXiv:2009.08453* (2020).

[37] Dimitrios Stamoulis, Ruizhou Ding, Di Wang, Dimitrios Lymberopoulos, Bodhi Priyantha, Jie Liu, and Diana Marculescu. 2019. Single-Path NAS: Designing Hardware-Efficient ConvNets in Less Than 4 Hours. In *ECML PKDD (Lecture Notes in Computer Science, Vol. 11907)*, Ulf Brefeld, Élisa Fromont, Andreas Hotho, Arno J. Knobbe, Marloes H. Maathuis, and Céline Robardet (Eds.). Springer, 481–497.

[38] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *The IEEE Conference on Computer Vision and Pattern Recognition*. 2818–2826.

[39] Shitao Tang, Litong Feng, Wenqi Shao, Zhanghui Kuang, Wei Zhang, and Yimin Chen. 2019. Learning efficient detector with semi-supervised adaptive distillation. *arXiv preprint arXiv:1901.00366* (2019).

[40] Shitao Tang, Litong Feng, Wenqi Shao, Zhanghui Kuang, Wayne Zhang, and Zheng Lu. 2019. Learning Efficient Detector with Semi-supervised Adaptive Distillation. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*. BMVA Press, 215.

[41] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive Representation Distillation. In *International Conference on Learning Representations*.

[42] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, Vol. 139. 10347–10357.

[43] Frederick Tung and Greg Mori. 2019. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*. 1365–1374.

[44] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. 2020. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence* (2020).

[45] Tiancheng Wen, Shenqi Lai, and Xueming Qian. 2019. Preparing lessons: Improve knowledge distillation with better supervision. *arXiv preprint arXiv:1911.07471* (2019).

[46] Ross Wightman. 2019. PyTorch Image Models. https://github.com/rwightman/pytorch-image-models. https://doi.org/10.5281/zenodo.4414861

[47] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. 2017. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4133–4141.

[48] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. 2021. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986* (2021).

[49] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. 2020. Revisiting Knowledge Distillation via Label Smoothing Regularization. In *The IEEE Conference on Computer Vision and Pattern Recognition*. 3903–3911.

[50] Sergey Zagoruyko and Nikos Komodakis. 2017. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. In *International Conference on Learning Representations*.

[51] Youcai Zhang, Zhonghao Lan, Yuchen Dai, Fangao Zeng, Yan Bai, Jie Chang, and Yichen Wei. 2020. Prime-Aware Adaptive Distillation. In *European Conference on Computer Vision (Lecture Notes in Computer Science, Vol. 12364)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer, 658–674.

[52] Zhilu Zhang and Mert R. Sabuncu. 2020. Self-Distillation as Instance-Specific Label Smoothing. In *Advances in Neural Information Processing Systems*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.).

[53] Helong Zhou, Liangchen Song, Jiajie Chen, Ye Zhou, Guoli Wang, Junsong Yuan, and Qian Zhang. 2020. Rethinking Soft Labels for Knowledge Distillation: A Bias–Variance Tradeoff Perspective. In *International Conference on Learning Representations*.