

A Fast Knowledge Distillation Framework for Visual Recognition

Zhiqiang Shen Eric Xing
CMU & MBZUAI

zhiqians@andrew.cmu.edu epxing@cs.cmu.edu

Abstract

While Knowledge Distillation (KD) has been recognized as a useful tool in many visual tasks, such as supervised classification and self-supervised representation learning, the main drawback of a vanilla KD framework is its mechanism, which consumes the majority of the computational overhead on forwarding through the giant teacher networks, making the entire learning procedure inefficient and costly. ReLabel [52], a recently proposed solution, suggests creating a label map for the entire image. During training, it receives the cropped region-level label by RoI aligning on a pre-generated entire label map, allowing for efficient supervision generation without having to pass through the teachers many times. However, as the KD teachers are from conventional multi-crop training, there are various mismatches between the global label-map and region-level label in this technique, resulting in performance deterioration. In this study, we present a Fast Knowledge Distillation (FKD) framework that replicates the distillation training phase and generates soft labels using the multi-crop KD approach, while training faster than ReLabel since no post-processes such as RoI align and softmax operations are used. When conducting multi-crop in the same image for data loading, our FKD is even more efficient than the traditional image classification framework. On ImageNet-1K, we obtain 79.8% with ResNet-50, outperforming ReLabel by ~1.0% while being faster. On the self-supervised learning task, we also show that FKD has an efficiency advantage. Our project page is [here](#), source code and models are available at: <https://github.com/szq0214/FKD>.

1. Introduction

Knowledge Distillation (KD) [16] has been a widely used technique in various visual domains, such as the supervised recognition [2, 24, 30, 35, 47, 48] and self-supervised representation learning [4, 10, 33]. The mechanism of knowledge distillation is to force the student to imitate the output of a teacher network or ensemble teachers, as well as converging on the ground-truth labels. Given the param-

Table 1. A feature-by-feature comparison between ReLabel [52] and our FKD framework on various elements and properties.

Method	Generating label	Label storage	Info. loss	Training
Vanilla KD	Implicit	None	No	Slow
ReLabel [52]	Fast	Efficient	Yes	Fast
FKD (Ours)	Slow	Efficient	No	Faster

eters θ of the target student at iteration (t), we can learn the next iteration parameters $\theta^{(t+1)}$ by minimizing the following objective which contains two terms:

$$\theta_{\text{student}}^{(t+1)} = \arg \min_{\theta \in \Theta} \frac{1}{N} \sum_{n=1}^N (1 - \lambda) \mathcal{H}(\mathbf{y}_n, \mathbf{S}_\theta(\mathbf{x}_n)) + \lambda \mathcal{H}(\mathbf{T}^{(t)}(\mathbf{x}_n), \mathbf{S}_\theta(\mathbf{x}_n)) \quad (1)$$

where \mathbf{y}_n is the ground-truth label for n -th sample. $\mathbf{T}^{(t)}$ is the teacher's output at iteration (t) and $\mathbf{S}_\theta(\mathbf{x}_n)$ is the student's prediction for the input sample \mathbf{x}_n . \mathcal{H} is the cross-entropy loss function. λ is coefficient for balancing the two objectives. The first term aims to minimize the entropy between one-hot ground-truth label and student's prediction while the second term is to minimize between teacher and student's predictions. The teacher \mathbf{T} can be pre-trained in either a supervised or self-supervised manner. Many literature [2, 35, 36, 52] have empirically shown that the first term of true hard label in Eq. 1 is not required on larger-scale datasets like ImageNet [8] under the circumstance that the teacher or ensembled teachers are accurate enough. In this work, we simply minimize the soft predictions between teacher and student models for the fast distillation design.

The fundamental disadvantage in such a paradigm, according to KD's definition, is that a considerable proportion of computing resources is consumed on passing training data through large teacher networks to produce the supervision $\mathbf{T}^{(t)}$ in each iteration, rather than updating or training the target student parameters. Intuitively, the forward propagation through teachers can be shared across epochs since the parameters of them are frozen for the entire training. Considering that the vanilla distillation framework itself is basically inefficient, how to reduce or share the forward computing of teacher networks across different epochs becomes the core for accelerating KD frameworks. A natural solution to overcome this drawback is to generate one probability vector as the soft label of input data correspond-

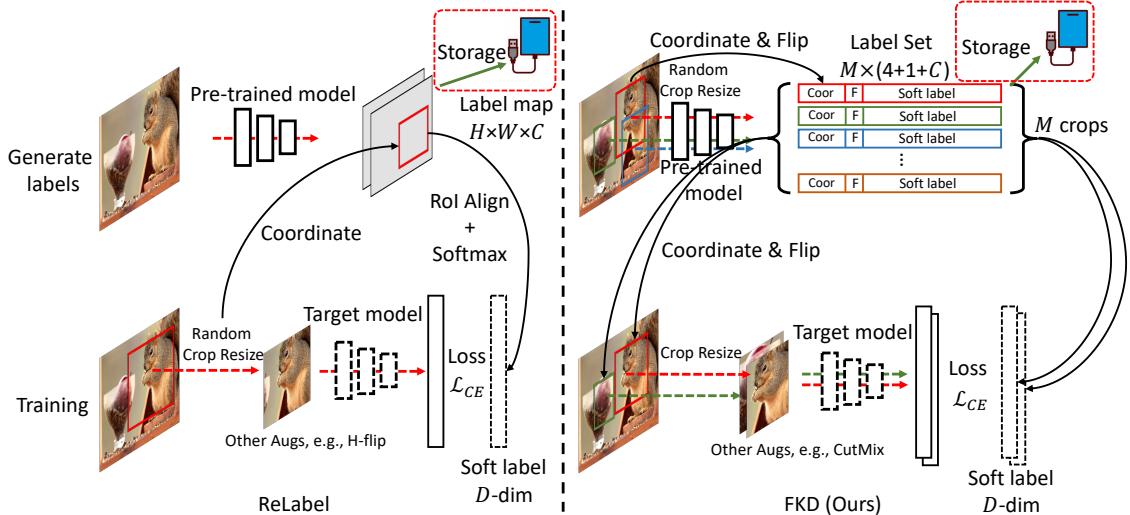


Figure 1. Comparison of ReLabel [52] and our fast knowledge distillation (FKD) framework.

ing to each image in advance, and then reuse them circularly for different training epochs. However, in modern network training, we usually impose various data augmentation strategies, particularly the random crop technique, which causes the inconsistency in which the simple global-level soft vector for the entire image can no longer accurately reflect the true probability distribution of the local input region after these augmentations.

To address the data augmentation, specially random-crop caused inconsistency issue in generating one global vector to the region-level input, while preserving the advantage of soft label property, ReLabel [52] is proposed to store the global label map annotations from a pre-trained strong teacher for reutilization by RoI align [13] without passing through the teacher networks repeatedly. Fig. 1 (left) shows a full explanation of this mechanism. However, due to the different input processes on teachers, this strategy is essentially not equivalent to the vanilla KD procedure. The mismatches are primarily due to two factors: **(i)** the teacher network is usually trained with a random-crop-resize scheme, whereas in ReLabel, the global label map is obtained by feeding into the global image, which cannot exactly reflect the soft distribution as distillation process whose random-crop-resize operation is employed in the input space; **(ii)** RoI align cannot guarantee the distribution is completely identical to that from the teachers’ forwarding.

In this work, we introduce a Fast Knowledge Distillation (FKD) framework to overcome the mismatching drawback and further avoid information loss on soft labels. Our strategy is straightforward: As shown in Fig 1 (right), in the label generation phase, we directly store the soft probability from multiple random-crops into the label files, together with the coordinates and other data augmentation status like flipping. During training, we assign these stored coordinates back to the input image to generate the crop-resized

input for passing through the networks, and computing the loss with the corresponding soft labels. The advantages of such a strategy are two folders: **(i)** Our region-based generating process and obtained soft label for each input region are identical to the vanilla KD’s output, implying that no information is lost during the label creation phase; **(ii)** Our training phase enjoys a faster pace since no post-process is required, such as RoI align, softmax, etc. We can further assign multiple regions from the same image in a *mini*-batch to facilitate the burden of data loading.

We demonstrate the advantages of our FKD in terms of accuracy and training speed on supervised and self-supervised learning tasks. In the supervised learning scheme, we compare the baseline ReLabel and vanilla KD (Oracle) from scratch across a variety of backbone network architectures, such as CNNs, vision transformers, and the competitive MEAL V2 framework with pre-trained initialization. Our FKD is $\sim 1\%$ higher and slightly faster than ReLabel on ImageNet-1K, and $3\sim 5\times$ faster than oracle KD with similar performance. On the self-supervised learning manner, we employ S²-BNN as the baseline for verifying the speed advantage of our proposed framework.

Our contributions of this work:

- We present a fast knowledge distillation (FKD) framework that achieves the same high level of performance as vanilla KD, while keeping the same training speed and efficiency as non-KD training without information loss.
- We reveal a discovery that in image classification frameworks, one image can be sampled with multiple crops within a *mini*-batch to facilitate data loading and speed up training, meanwhile without sacrificing performance.
- To prove the effectiveness and versatility of our approach, we demonstrate FKD on a variety of tasks and distillation frameworks, including supervised classification and self-supervised representation learning.

2. Related Work

Knowledge Distillation. The principle behind Knowledge Distillation [16] is that a student is encouraged to emulate or mimic the teachers’ prediction, which helps the student generalize better on unseen data. One core advantage of distillation is that the teacher can provide softened distribution which contains richer information about the input data compared to the traditional one-hot labels, especially when the data augmentation such as random cropping is used on the input space. Distillation can avoid incorrect labels by predicting them from the strong teachers in each iteration, which reflects the real situation of the transformed input data. Conventionally, we can impose a temperature on the logits to re-scale the output distributions from teacher and student models to amplify the inter-class relationship on supervisions and allow for improved distillation. Recently, many variants and extensions are proposed [6, 18, 24–26, 35, 37, 44, 47, 49, 54], such as employing internal feature representations [30], adversarial training with discriminators [32], transfer flow [48], contrastive distillation [40], patient and consistent [2] etc. For the broader overviews of related methods for knowledge distillation, please refer to [12, 43].

Efficient Knowledge Distillation. Improving training efficiency for knowledge distillation is crucial for pushing this technique to a wider usage scope in real-world applications. Previous efforts on this direction are generally not sufficient. ReLabel [52] is a recently proposed solution that addresses this inefficient issue of KD surpassingly. In particular, it generates the global label map for the strong teacher and then reuses them through RoI align across different epochs. Our proposed FKD approach in this paper lies in an essentially different consideration and solution. We consider the property of vanilla KD to generate the randomly cropped region-level soft labels from the strong teachers and store them in advance, then reuse them by allocating them to different epochs in training. Our approach enjoys the same accuracy as vanilla KD and same training speed as regular non-KD classification frameworks, making it superior than ReLabel in both performance and training speed.

3. Approach

In this section, we begin by introducing some observations and properties on ReLabel’s global-level soft label and FKD’s region-level soft label distributions. Then, we present the detailed workflow of our FKD framework and elaborately discuss the generated label quality, training speed and the applicability on supervised and self-supervised learning. Finally, we analyze the strategies of label compression and storage for practical usage.

Preliminaries: Limitations of Previous Solution

The mechanism of ReLabel through RoI align operation (an approximation solution) is naturally different from the

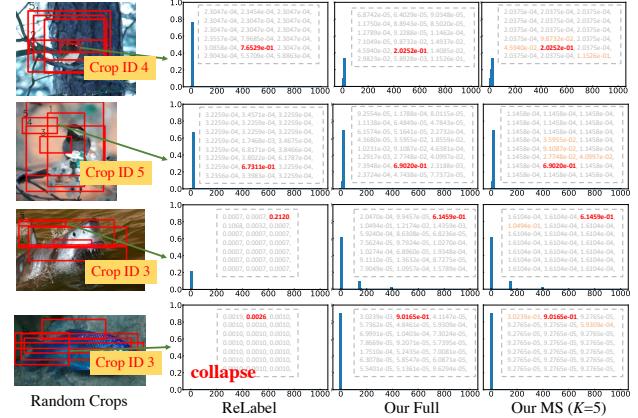


Figure 2. Illustration of label distributions of ReLabel [52], our FKD full label and our quantized label (Top-5). “MS” denotes the marginal smoothed labels, more details can be referred in Sec. 3.5. Grey numbers in each block are the corresponding partial probabilities/labels (limited by space).

vanilla KD when generating region-level soft labels. In Fig. 2, we visualize the region-level label distributions of ReLabel and FKD on ImageNet-1K, and several empirical observations are noticed: **(i)** ReLabel is more confident in many cases of the regions, so the soft information is weaker than our FKD. We conjecture this is because ReLabel feeds the global images into the network instead of local regions, which makes the generated global label map encode more global category information and ignores the backgrounds, as shown in Fig. 2 (row 1). Though sometimes the maximal probabilities are similar between ReLabel and FKD, FKD still contains more informative subordinate probabilities in the label distribution, as shown in Fig. 2 (row 2); **(ii)** for some outlier regions, our strategy is substantially more robust than ReLabel, such as the loose bounding boxes of objects, partial object, etc., as shown in Fig. 2 (row 3); **(iii)** In some particular circumstance, ReLabel is unexpectedly collapsed with nearly uniform distribution, while our FKD still works well, as shown in the bottom row of Fig. 2.

Moreover, there are existing mismatches between the soft label from ReLabel and oracle teacher prediction in KD when employing more data augmentations such as Flip, Color jittering, etc., since these augmentations are randomly applied during training. In ReLabel design, we cannot take them into account and prepare in advance when generating the global label map. In contrast, FKD is adequate to handle this situation: it is effortless to involve extra augmentations and record all information (ratio, degree, coefficient, etc.) for individual region from same or different images, and generate corresponding soft label by feeding the transformed image regions into the pre-trained teacher networks.

3.1. Fast Knowledge Distillation

In a conventional visual training system, the bottleneck is usually from the network passing and data loader, while in

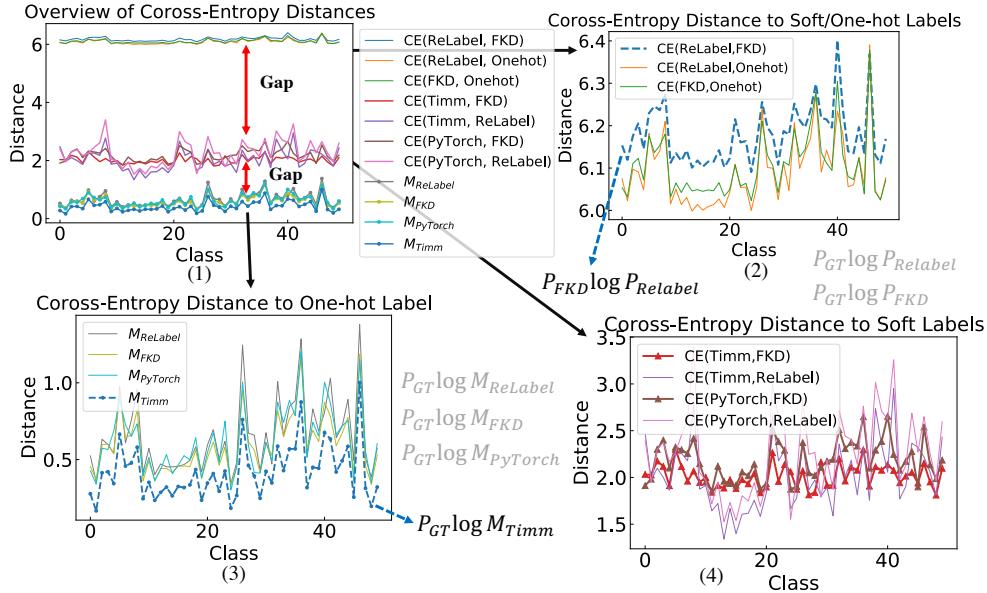


Figure 3. Entropy distance analysis between different pairs of soft/one-hot labels and different labels trained model predictions. (1) is the overall distance visualization. (2), (3), (4) represent each detailed group in (1). We illustrate the first 50 classes in ImageNet-1K dataset.

a distillation framework, besides these computational consumptions, giant teachers have been the biggest burden for training. Our FKD aims to solve this intractable drawback.

Label Generation Phase. Following the regular random-crop resize training strategy, we randomly crop M regions from one image and employ other augmentations like flipping on them, then input these regions into the teachers to generate the corresponding soft label vectors \mathbf{P}_i , i.e., $\mathbf{P}_i = \mathbf{T}(\mathbf{R}_i)$ where \mathbf{R} is the transformed region by transformations \mathcal{F} and \mathbf{T} is the pre-trained teacher network, i is the region index. We store all the region coordinates and augmentation hyper-parameters \mathcal{F} with the soft label \mathbf{P} for the following training phase, as shown in Fig. 1 (upper right). A detailed analysis of how to store these required values is provided in the following section.

Training Phase. In the training stage, instead of randomly generating crops as the conventional image classification strategy, we directly load the label file, and assign our stored crop coordinates and data augmentations for this particular image to prepare the transformed region-level inputs. The corresponding soft label will be used as the supervision of these regions for training. With the cross-entropy loss, the objective is: $\mathcal{L} = -\sum_i \mathbf{P}_i \log S_\theta(\mathbf{R}_i)$, where $S_\theta(\mathbf{R}_i)$ is the student’s prediction for the input region \mathbf{R}_i , θ is the parameter of the student model that we need to learn. The detailed training procedure is shown in Fig. 1 (bottom right).

3.2. Higher Label Quality

Distance Analysis. We analyze the quality of various formulations of labels through the entropy distance with measures on their mutual cross-entropy matrix. We consider three types of labels: (1) human-annotated one-hot label,

ReLabel, and our FKD. We also calculate the distance on the predictions of four pre-trained models with different accuracies, including: PyTorch pre-trained model (weakest), Timm pre-trained model [46] (strongest), ReLabel trained model and FKD trained model. An overview of our illustration is shown in Fig. 3. The upper curves, as well in (2), are averaged cross-entropy across 50 classes of (ReLabel→FKD), (ReLabel→One-hot) and (FKD→One-hot). Here, we derive an important observation:

$$(\mathcal{D}_{R \rightarrow F}^{CE} = -P_{FKD} \log P_{ReLabel}) > (\mathcal{D}_{R \rightarrow O}^{CE} \text{ OR } \mathcal{D}_{F \rightarrow O}^{CE}) \quad (2)$$

where $\mathcal{D}_{R \rightarrow F}^{CE}$ is the cross-entropy value of ReLabel → FKD. Essentially, FKD soft label can be regarded as the oracle KD label and $\mathcal{D}_{R \rightarrow F}^{CE}$ is the distance to such “KD ground truth”. From Fig. 3 (2) we can see its distance is even large than ReLabel and FKD to the one-hot label. Since ReLabel (global-map soft label) and FKD (region-level soft label) are greatly discrepant from the one-hot hard label, the gap between ReLabel and FKD (“KD ground truth”) is fairly significant and considerable. If we shift attention to the curves of $\mathcal{D}_{R \rightarrow O}^{CE}$ and $\mathcal{D}_{F \rightarrow O}^{CE}$, they are highly aligned across different classes with similar values. In some particular classes, $\mathcal{D}_{F \rightarrow O}^{CE}$ are slightly larger. This is sensible as one-hot label is basically not the “optimal label” we desired.

In the bottom group, i.e., Fig. 3 (3), the entropy values are comparatively small. This is because they are from the pre-trained models and they have the decent performance under the criterion metric of one-hot label. Among them, M_{Timm} has the minimal cross-entropy to the one-hot label, this is expected since the timm model is optimized thoroughly to fit the one-hot label with the highest accuracy.

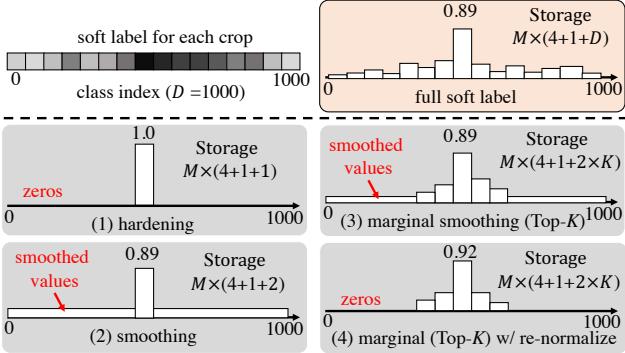


Figure 4. Different label compression strategies and storage analysis for our fast knowledge distillation (FKD) framework.

In Fig. 3 (4), $\mathcal{D}_{Timm \rightarrow F}^{CE}$ and $\mathcal{D}_{PT \rightarrow F}^{CE}$ lie in the middle of $\mathcal{D}_{Timm \rightarrow R}^{CE}$ and $\mathcal{D}_{PT \rightarrow R}^{CE}$ with smaller variances. This reflects that FKD is more stable to the pre-trained models.

3.3. Faster Training Speed

Multi-crop sampling within a *mini-batch*. As illustrated in Fig. 1 (right), we can use multiple crops in the same image to facilitate loading image and label files. Intuitively, it will reduce the diversity of training samples in a *mini-batch* since some of samples are from the same image. However, our experimental results indicate that it will not hurt the model performance, in contrast, it even boosts the accuracy when the number of crops from the same image is within a reasonable range (e.g., 4~8).

Serrated learning rate scheduler. Since FKD samples multiple crops ($\#crop$) from one image, when iterating over the entire dataset once, we actually train the dataset $\#crop$ epochs with the same learning rate. It has no effect while using milestone/step lr scheduler, but it will change the lr curve to be serrated if applying continuous *coseine* or *linear* learning rate strategies. This is also the potential reason that multi-crop training can improve the accuracy.

Training Time Analysis: 1. Data Load

Data loading strategy in FKD is efficient. For instance, when training with a *mini-batch* of 256, traditional image classification framework requires to load 256 images and ReLabel will load 256 images + 256 label files, while in our method, FKD only needs to load $\frac{256}{\#crop}$ images + $\frac{256}{\#crop}$ label files, even faster than traditional training if we choose a slightly larger value for $\#crop$ (when $\#crop > 2$)¹.

2. Label Preparation

We assign $\#crop$ regions in an image to the current *mini-batch* for training. Since we store the label probability after *softmax* (in supervised learning), we can use assigned soft labels for the *mini-batch* samples directly without any post-process. This assignment is fast and efficient in implementation with a *randperm* function in PyTorch [27]. If the label

¹We assume that loading each image and label file will consume the similar time by CPUs.

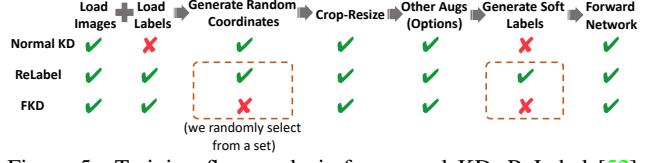


Figure 5. Training flow analysis for normal KD, ReLabel [52] and our fast knowledge distillation (FKD) framework. Maroon dashed boxes indicate that the processes are required by ReLabel only while not existing in our FKD. Note that “generate soft labels” indicates RoI align + softmax in ReLabel. We both have the recovering process from the compressed label to full soft label as discussed in Sec. 3.3.

is compressed using the following strategies, we will operate with an additional simple recovering process (as shown in Fig. 4) to obtain D -way soft label distributions. Note that ReLabel also has this process so the time consumption on this part will be similar to ReLabel. A detailed workflow and item-by-item comparison is shown in Fig. 5.

3.4. Training Self-supervised Model with Supervised Scheme

In this section, we introduce how to apply our FKD for extending to the self-supervised learning (SSL) with faster training speed, comparing to the widely-used Siamese SSL frameworks. The label generation (from the self-supervised strong teachers), label preparation and training procedure are similar to the supervised scheme. However, we *keep the projection head in original SSL teachers and store the soft labels before softmax* for operating temperature².

3.5. Label Compression and Storage Analysis

We consider and introduce the following four strategies for compressing soft label for storage, an elaborated comparison of them can be referred to Table 2.

- **Hardening.** In hardening quantization strategy, the hard label \mathbf{Y}_H is generated using the index of the maximum logits from the teacher predictions of regions. In general, label hardening is the one-hot label with correction by strong teacher models in region-level space.

$$\mathbf{Y}_H = \operatorname{argmax}_c z_{\text{FKD}}(c) \quad (3)$$

where z_{FKD} is the logits for each randomly cropped region produced by our FKD process.

- **Smoothing.** Smoothing quantization replaces one-hot hard label \mathbf{Y}_H with a mixture of soft \mathbf{y}_c and a uniform distribution same as label smoothing [38]:

$$y_c^S = \begin{cases} p_c & \text{if } c = \text{hardening label}, \\ (1 - p_c)/(C - 1) & \text{otherwise.} \end{cases} \quad (4)$$

where p_c is the probability after *softmax* at c -th class and C is the number of total classes. $(1 - p_c)/(C - 1)$ a

²The temperature τ is applied on the *logits* before the *softmax* operation for self-supervised distillation.

Table 2. A detailed comparison of different label quantization/compression strategies on ImageNet-1K. M is the number of crops within an image and here we choose 200 crops as an example to calculate the space consumption. N_{image} is the number of images, i.e., 1.2M for ImageNet-1K. S_{LM} is the size of label map. C_{class} is the number of classes. D_{DA} is the parameter dimension of data augmentations to store.

	ReLabel (Full) [52]	ReLabel (Top-5) [52]	Full	Hard	Smoothing	M Re-Norm ($K=5$)	MS ($K=5$)	MS ($K=10$)
Calculation	$N_{\text{image}} \times S_{\text{LM}} \times C_{\text{class}}$	$N_{\text{image}} \times S_{\text{LM}} \times 2C_{\text{Top-5}}$	$N_{\text{image}} \times (C_{\text{class}} + D_{\text{DA}})$	$N_{\text{image}} \times (1 + D_{\text{DA}})$	$N_{\text{image}} \times (2 + D_{\text{DA}})$	$N_{\text{image}} \times (2K + D_{\text{DA}})$	$N_{\text{image}} \times (2K + D_{\text{DA}})$	$N_{\text{image}} \times (2K + D_{\text{DA}})$
Dim. of Soft Label + Coordinate & Flip	$15 \times 15 \times 1,000$	$15 \times 15 \times 10$	$M \times 1,000$	$M \times 1$	$M \times 2$	$M \times 6$	$M \times 10$	$M \times 20$
Real Cons. on Disk	–	–	$M \times 1,005$	$M \times 6$	$M \times 7$	$M \times 15$	$M \times 15$	$M \times 25$
	$\sim 1\text{TB}$	10GB	$\sim 0.9\text{TB}$	5.3GB	6.2GB	13.3GB	13.3GB	22.2GB

small value for flattening the one-hot labels. $y_c^S \in Y_S$ is the smoothed label at c -th class.

- **Marginal Smoothing with Top- K (MS).** Marginal smoothing quantization reserves more soft information (Top- K) of teacher prediction than the single smoothing label Y_S :

$$y_c^{\text{MS}} = \begin{cases} p_c & \text{if } c \in \{\text{Top}-K\}, \\ \frac{1 - \sum_{c \in \{\text{Top}-K\}} p_c}{C-K} & \text{otherwise.} \end{cases} \quad (5)$$

where $y_c^{\text{MS}} \in Y_{\text{MS}}$ is the marginally smoothed label at c -th class.

- **Marginal Re-Norm with Top- K (MR).** Marginal re-normalization will re-normalize Top- K predictions to $\sum_{c \in \{\text{Top}-K\}} p_c = 1$ and maintain other logits to be zero (this strategy is spiritually similar to ReLabel [52] but slightly different in implementation as its input is logits before softmax so it used *softmax* while we use *normalize*, resulting in that our values outside Top- K remain zero.):

$$y_c^{\text{M}} = \begin{cases} p_c & \text{if } c \in \{\text{Top}-K\}, \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

$$y_c^{\text{MR}} = \text{Normalize}(y_c^{\text{M}}) = \frac{y_c^{\text{M}}}{\sum_{c=1}^C (y_c^{\text{M}})} \quad (7)$$

where $y_c^{\text{MR}} \in Y_{\text{MR}}$ is the re-normalized label at c -th class.

4. Experiments

Experimental Settings and Datasets. Detailed lists of our hyper-parameter choices are shown in Table 10, Table 11 of Appendix. For the transparency and reproducibility of our framework, unless noted otherwise, we did not involve extra data augmentations (beyond the basic random crop and random horizontal flip) such as RandomAug [7], MixUp [53], CutMix [51], etc., in all of our experiments. Except for experiments on MEAL V2, we use *EfficientNet-L2-ns-475* [39, 47] as the teacher model, we also tried weaker teacher but the performance in our experiment is slightly worse. For MEAL V2, we follow its original design by using *SENet154 + ResNet152_v1s* (gluon version) ensemble as the soft label.

ImageNet-1K [8] is used for the supervised classification and self-supervised representation learning. COCO [21] is used for the transfer learning experiments in this work.

Table 3. Comparison between ReLabel [52] and our FKD on ImageNet-1K. “◊” denotes our training following the same protocol in Table 10 w/o distillation. Models are trained from scratch.

Method	Network	Top-1 (%)	Top-5 (%)	Training Time
Vanilla◊	ResNet-50	78.1	94.0	1.0
ReLabel [52]	ResNet-50	78.9	–	$\uparrow 0.5\%$ [52]
FKD (Ours)	ResNet-50	79.8^{+0.9}	94.6	$\downarrow 0.5\%$
Vanilla◊	ResNet-101	79.7	94.6	1.0
ReLabel [52]	ResNet-101	80.7	–	$\uparrow 0.5\%$ [52]
FKD (Ours)	ResNet-101	81.7^{+1.0}	95.6	$\downarrow 0.5\%$

Table 4. Comparison of MEAL V2 [36] and our FKD on ImageNet-1K. “w/ FKD” denotes the model is trained using the same protocol as original MEAL V2, i.e., all the same hyper-parameters. “◊” represents the training using *cosine lr* and $1.5 \times$ epochs. Models are trained from the pre-trained initialization.

Method	Network	#Params	Top-1	Top-5	Speedup
MEAL V2 [36]	ResNet-50	25.6M	80.67	95.09	1.0
MEAL V2 w/ FKD	ResNet-50	25.6M	80.70	95.13	0.3×
MEAL V2 w/ ♦FKD	ResNet-50	25.6M	80.91	95.39	0.5×
MEAL V2 [36]	MobileNet V3-S0.75	2.04M	67.60	87.23	1.0
MEAL V2 w/ ♦FKD	MobileNet V3-S0.75	2.04M	67.83	87.35	0.4×
MEAL V2 [36]	MobileNet V3-S1.0	2.54M	69.65	88.71	1.0
MEAL V2 w/ ♦FKD	MobileNet V3-S1.0	2.54M	69.94	88.82	0.4×

Network Architectures. Experiments are conducted on Convolutional Neural Networks [19], such as ResNet [14], MobileNet [17] and Vision Transformers [9, 42], such as DeiT [41], SReT [34]. For binary backbone, we use ReActNet [22] in the self-supervised experiments.

Learning Schemes. We consider three training manners in vision tasks: (i) conventional supervised training from scratch; (ii) supervised fine-tuning from pre-trained parameters; and (iii) self-supervised distillation from scratch.

Baseline Knowledge Distillation Methods.

- **ReLabel [52] (Label Map Distillation).** ReLabel used the pre-generated global label maps from the pre-trained teacher for reducing the cost on the teacher branch when conducting distillation.
- **MEAL V2 [36] (Fine-tuning Distillation).** MEAL V2 proposed to distill student network from the pre-trained parameters³ and giant teacher ensemble for fast convergence and better accuracy.
- **FunMatch [52] (Oracle Distillation).** FunMatch is a standard knowledge distillation framework with strong teacher models and augmentations. We consider it as the strong baseline approach for efficient KD when using the same or similar teacher supervisors.

³The pre-trained parameter is from *timm* [45] with version $<= 0.4.12$.

Table 5. FKD with supervised Vision Transformer [9] variants (224×224 input size) on ImageNet-1K. Models are trained from scratch.

Method	Network	Epochs	#Params (M)	FLOPs (B)	Extra Data Aug.	Top-1 (%)	Speedup
DeiT [41] w/o KD	ViT-T	300	5.7	1.3	MixUp+CutMix+RA	72.2	–
DeiT [41] w/ KD	ViT-T	300	5.7	1.3	MixUp+CutMix+RA	74.5	1.0
ViT [9] (Vanilla)	ViT-T	300	5.7	1.3	None	68.7 [15]	–
ViT w/ FKD (Ours)	ViT-T	300	5.7	1.3	None	75.2	0.15×
SReT [34] w/o KD	SReT-LT	300	5.0	1.2	MixUp+CutMix+RA	76.7	–
SReT [34] w/ KD	SReT-LT	300	5.0	1.2	MixUp+CutMix+RA	77.7	1.0
SReT [34] (Vanilla)	SReT-LT	300	5.0	1.2	None	–	–
SReT w/ FKD (Ours)	SReT-LT	300	5.0	1.2	None	78.7	0.14×

Table 6. Ablation results (Top-1) on ImageNet-1K of different label quantization strategies. $m = 8$ is used in this ablation.

Method	Network	Full	Hard	Smoothing	Marginal Re-Norm ($K=5$)	Marginal Smoothing ($K=5$)	Marginal Smoothing ($K=10$)
MEAL V2 w/ FKD	ResNet-50	80.65	80.20	80.23	80.40	80.58	80.52
FKD (from scratch)	ResNet-50	79.48	79.09	79.37	79.23	79.51	79.44

Table 7. Ablation results (Top-1) on ImageNet-1K of different numbers (m) of cropping regions within an image used in a *mini-batch*.

Method	Network	$m = 1$	$m = 2$	$m = 4$	$m = 8$	$m = 16$	$m = 32$
Vanilla	ResNet-50	77.18	77.91	78.14	77.89	75.89	70.09
MEAL V2 w/ FKD	ResNet-50	80.67	80.70	80.66	80.58	80.36	80.17
FKD (from scratch)	ResNet-50	79.59	79.62	79.76	79.51	78.12	74.61

▷ S²-BNN [52] (Self-supervised Distillation). A distillation solution for self-supervised learning task. The teacher is pre-learned from the self-supervised learning methods, such as MoCo V2 [5], SwAV [3], etc.

4.1. Supervised Learning

CNNs.

(i) **ReLabel** [52]. The comparison with ReLabel is shown in Table 3, using the training settings introduced in our Appendix, which is nearly the same as ReLabel, our accuracies on ResNet-50/101 both outperform ReLabel by $\sim 1\%$ with slightly faster training speed. These significant and consistent improvements of FKD show great potential for practical usages in real-world applications.

(ii) **MEAL V2** [36]. We use FKD framework to train the MEAL V2 models. The results are shown in Tabel 4. When employing the same hyper-parameters and teacher networks, FKD can speed up $2\sim 4\times$ for MEAL V2 **without compromising accuracy**. Using *cosine lr* and more epochs in training further improves the accuracy.

(iii) **FunMatch** [2] (Oracle). We consider FunMatch as the oracle/strong KD baseline, our plain FKD achieves 79.8% w/o extra augmentations, which is 0.7% lower than FunMatch (80.5%). The result of FKD with more complex optimizers [23, 50] and more augmentations [51, 53] (similar to FunMatch) will be presented after being explored.

Vision Transformers.

(i) **ViT/DeiT** [9, 41]. The results are shown in Tabel 5 of first group. Our non-extra augmentation result (75.2%) using ViT-T backbone is better than DeiT-T with distillation (74.5%), while we only require **0.15×** resources than DeiT distillation protocol in training.

(ii) **SReT** [34]. We also conduct our FKD using SReT-LT, our result (78.7%) is consistently better than its original

KD design (77.7%) with fewer training resources.

Ablations: (i) Effects of Crop Number in Each Image.

We explore the effect of different numbers of crops sampled from the same image within a *mini-batch* to the final performance. For the conventional data preparation strategy, on each image we solely sample one crop for a *mini-batch* to train the model. Here, we evaluate the m from 1 crop to 32 crops as shown in Table 7. Surprisingly, using a few crops from the same image leads to better performance than the single crop solution with a non-negligible margin, especially on the traditional image classification system. This indicates that the internal diversity of samples in a *mini-batch* has a limit for tolerance, properly reducing such diversity can boost the accuracy, while we can also observe that after $m > 8$, the performance decreases substantially, thus the diversity is basically still critical for learning good status of the model. Nevertheless, this is a good observation for us to speedup data loading in our FKD framework.

(ii) **Different Label Compression Strategies.** We evaluate the performance for different label compression strategies. We use $m=8$ for this ablation and the results are shown in Tabel 6. On MEAL V2 w/ FKD, we obtain the highest accuracy 80.65% when using the full soft labels, while on the pure FKD, the best performance is from *Marginal Smoothing ($K=5$)* with 79.51%. Increasing K both decrease the accuracies in these two scenarios, we conjecture that larger K will involve more noise or unnecessary minor information on the soft labels. While, they are still better than the *Hard* and *Smoothing* strategies.

4.2. Self-Supervised Learning

S²-BNN [33] is a pure distillation-based framework for self-supervised learning, thus the proposed FKD approach is sufficient to train S²-BNN [33] smoothly and more efficiently. We employ SwAV [3] and MoCo V2 [5] pre-trained

Table 8. Linear evaluation results of FKD with self-supervised Binary CNN (ReActNet [22]), Real-valued CNN (ResNet-50 [14]).

Method	Network	Teacher	#Dim for Distilling	Training Epochs	Top-1 (%)	Speedup
<i>Pure Distillation Based Scheme (FKD can speedup training by more than 3× with the same (similar) performance.)</i>						
S ² -BNN [33]	ReActNet	MoCo V2-800ep	128	200	61.5	1.0
FKD	ReActNet	MoCo V2-800ep	128	200	61.7	0.4×
S ² -BNN [33]	ResNet-50	SwAV/RN50-w4	3000	100	68.7	1.0
FKD	ResNet-50	SwAV/RN50-w4	3000	100	68.8	0.3×

Table 9. Comparison of transfer learning performance with ReLabel [52] on detection and instance segmentation tasks. The training and evaluation are conducted on COCO dataset [21].

Method	Network	Faster RCNN w/ FPN	Mask-RCNN w/ FPN	
		bbox AP	bbox AP	mask AP
Regular Baseline	ResNet-50	37.7	38.5	34.7
ReLabel	ResNet-50	38.2	39.1	35.2
FKD	ResNet-50	38.7	39.7	35.9

models as the teacher networks. Consider the more flattening distribution from the SSL learned teachers than the supervised teacher predictions (indicating that the subordinate classes from SSL trained teachers also carry crucial information), we still use the full soft label for now, and label compression strategies on SSL task will be verified further. We employ ReActNet [22] and ResNet-50 [14] as the target/student backbones in these experiments. The results are shown in Table 8, our FKD trained models achieve slightly better performance than S²-BNN with roughly 3× acceleration in training since we only use a single branch of network and no explicitly teacher forwarding existing. The slight boosts are from our liter data augmentation for FKD when producing SSL soft labels. This is interesting in the FKD equipped SSL methods, i.e., data augmentation strategies for distillation-based SSL, and is worth exploring further.

4.3. Transfer Learning

Here, we further examine whether the FKD obtained improvements on ImageNet-1K can be transferred to various downstream tasks. As shown in Tabel 9, we present the results of object detection and instance segmentation tasks on COCO dataset [21] with models pretrained on ImageNet-1K with FKD. We also employ Faster RCNN [29] and Mask RCNN [13] with FPN [20] following ReLabel [52]. Over the regular baseline and ReLabel, our FKD pre-trained parameters show constant gains on the downstream tasks.

4.4. Visualization, Analysis and Discussion

To investigate the learned differences of information between ReLabel and FKD, we depict the intermediate attention maps using gradient-based localization [31]. There are three important observations that align our aforementioned analyses in Fig. 6.

(i) FKD’s predictions are less confident than ReLabel with more surrounding context; This is reasonable since in random-crop training, many crops are basically back-

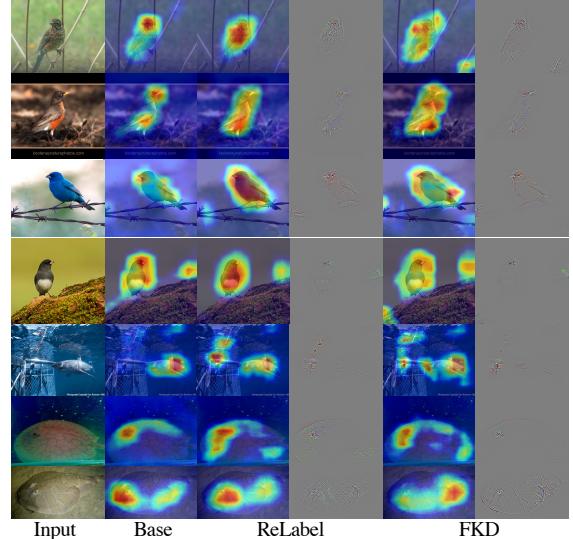


Figure 6. Visualizations of learned attention map using Grad-CAM [11, 31]. “Base” indicates the pre-trained PyTorch model. In each group of ReLabel and FKD, left is *Grad-CAM* and right is *Guided Backprop*.

grounds (context), the soft predicted label from the teacher model might be completely different from the ground-truth one-hot label and the training mechanism of FKD can leverage the additional information from context.

(ii) FKD’s attention maps have a larger active area on the object regions, which indicates that FKD trained model utilizes more cues for prediction and also captures more subtle and fine-grained information. However, it is interesting to see that the *guided backprop* is more focusing than ReLabel.

(iii) ReLabel’s attention is more aligned with PyTorch pre-trained model, while FKD’s results are substantially unique to them. It implies that FKD’s learned attention differs significantly from one-hot and global label map learned models.

5. Conclusion

Given its widespread use and superior performance in training compact and efficient networks, it is worthwhile investigating ways to increase the efficiency and speed of vanilla knowledge distillation. In this paper, we have presented a fast distillation framework through the pre-generated region-level soft label scheme. We have elaborately discussed the strategies of compressing soft label for

practical storage and their performance comparison. We identified the observation that the input training samples within a *mini*-batch can be sampled from the same input image to facilitate the overhead of data loading process. We exhibit the effectiveness and adaptability of our framework by demonstrating it on supervised image classification and self-supervised representation learning tasks.

References

- [1] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020. 11
- [2] Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. *arXiv preprint arXiv:2106.05237*, 2021. 1, 3, 7
- [3] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 7
- [4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 1
- [5] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 7
- [6] Inseop Chung, SeongUk Park, Jangho Kim, and Nojun Kwak. Feature-map-level online adversarial knowledge distillation. In *International Conference on Machine Learning*. PMLR, 2020. 3
- [7] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020. 6
- [8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 6
- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 6, 7, 11
- [10] Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. Seed: Self-supervised distillation for visual representation. In *ICLR*, 2021. 1
- [11] Jacob Gildenblat and contributors. Pytorch library for cam methods. <https://github.com/jacobgil/pytorch-cam>, 2021. 8
- [12] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021. 3
- [13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2, 8
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 8
- [15] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyu Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *International Conference on Computer Vision (ICCV)*, 2021. 7
- [16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1, 3
- [17] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 6
- [18] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017. 3
- [19] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 1995. 6
- [20] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 8
- [21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6, 8
- [22] Zechun Liu, Zhiqiang Shen, Marios Savvides, and Kwang-Ting Cheng. Reactnet: Towards precise binary neural network with generalized activation functions. In *European Conference on Computer Vision*, pages 143–159. Springer, 2020. 6, 8
- [23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 7
- [24] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? In *NeurIPS*, 2019. 1, 3
- [25] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, pages 582–597. IEEE, 2016. 3
- [26] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, pages 3967–3976, 2019. 3
- [27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019. 5

- [28] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, pages 5389–5400. PMLR, 2019. 11
- [29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 8
- [30] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 1, 3
- [31] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 8
- [32] Zhiqiang Shen, Zhankui He, and Xiangyang Xue. Meal: Multi-model ensemble via adversarial learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4886–4893, 2019. 3
- [33] Zhiqiang Shen, Zechun Liu, Jie Qin, Lei Huang, Kwang-Ting Cheng, and Marios Savvides. S2-bnn: Bridging the gap between self-supervised real and 1-bit neural networks via guided distribution calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2165–2174, 2021. 1, 7, 8
- [34] Zhiqiang Shen, Zechun Liu, and Eric Xing. Sliced recursive transformer. *arXiv preprint arXiv: 2111.05297*, 2021. 6, 7, 11
- [35] Zhiqiang Shen, Zechun Liu, Dejia Xu, Zitian Chen, Kwang-Ting Cheng, and Marios Savvides. Is label smoothing truly incompatible with knowledge distillation: An empirical study. In *ICLR*, 2021. 1, 3
- [36] Zhiqiang Shen and Marios Savvides. Meal v2: Boosting vanilla resnet-50 to 80%+ top-1 accuracy on imagenet without tricks. *arXiv preprint arXiv:2009.08453*, 2020. 1, 6, 7
- [37] Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A Alemi, and Andrew Gordon Wilson. Does knowledge distillation really work? *arXiv preprint arXiv:2106.05945*, 2021. 3
- [38] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 5
- [39] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 6
- [40] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2019. 3
- [41] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 6, 7, 11
- [42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 6
- [43] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 3
- [44] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018. 3
- [45] Ross Wightman. Pytorch image models. <https://github.com/rwightman/pytorch-image-models>, 2019. 6
- [46] Ross Wightman, Hugo Touvron, and Hervé Jégou. Resnet strikes back: An improved training procedure in timm. *arXiv preprint arXiv:2110.00476*, 2021. 4
- [47] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020. 1, 3, 6
- [48] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4133–4141, 2017. 1, 3
- [49] Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deep-inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8715–8724, 2020. 3
- [50] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017. 7
- [51] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019. 6, 7
- [52] Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, Junsuk Choe, and Sanghyuk Chun. Re-labeling imagenet: from single to multi-labels, from global to localized labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2340–2350, 2021. 1, 2, 3, 5, 6, 7, 8, 11
- [53] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 6, 7
- [54] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenguang Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *ICCV*, 2019. 3

Appendix

A. Training Details and Experimental Settings

Training details used in Table 3 of the main text. When comparing our FKD with ReLabel [52] (Table 3 of the main text), we use the training settings and hyper-parameters following Table 10, which is nearly the same as ReLabel [52] while without *Warmup* and *Color jittering*.

Table 10. Training hyper-parameters and details between ReLabel [52] and FKD used for the comparison in Table 3 of main text.

Method	ReLabel [52]	FKD
Teacher	EfficientNet-L2-ns-475	EfficientNet-L2-ns-475
Epoch	300	300
Batch size	1,024	1,024
Optimizer	SGD	SGD
Init. <i>lr</i>	0.1	0.1
<i>lr</i> scheduler	cosine	cosine
Weight decay	1e-4	1e-4
Random crop	Yes	Yes
Flipping	Yes	Yes
Warmup epochs	5	0
Color jittering	Yes	No

Training details used in Table 5 of the main text. When comparing our FKD with ViT [9]/DeiT [41]/SReT [34] (Table 5 of the main text), we use the training settings and hyper-parameters following Table 11.

Table 11. Training hyper-parameters and details for the comparison in Table 5 of the main text when employing ViT [9] and its variants as the backbone networks. Table is adapted from [41].

Method	ViT-B [9]	DeiT [41]/SReT [34]	FKD
Epoch	300	300	300
Batch size	4096	1024	1024
Optimizer	AdamW	AdamW	AdamW
Init. <i>lr</i>	0.003	0.001	0.002
<i>lr</i> scheduler	cosine	cosine	cosine
Weight decay	0.3	0.05	0.05
Warmup epochs	3.4	5	5
Label smoothing	None	0.1	None
Dropout	0.1	None	None
Stoch. Depth	None	0.1	0.1
Repeated Aug	None	Yes	None
Gradient Clip.	Yes	None	None
Rand Augment	None	9/0.5	None
Mixup prob.	None	0.8	None
Cutmix prob.	None	1.0	None
Erasing prob.	None	0.25	None

B. More Comparison and Results on ImageNet ReAL [1] and ImageNetV2 [28] Datasets

In this section, we provide more results on ImageNet ReAL [1] and ImageNetV2 [28] datasets. On ImageNetV2 [28], we verify our FKD models on three metrics “Top-Images”, “Matched Frequency”, and “Threshold

0.7” as ReLabel [52]. We conduct experiments on two network structures: ResNet-50 and ResNet-101. The results are shown in Table 12, we achieve consistent improvement over baseline ReLabel on both ResNet-50 and ResNet-101.

Table 12. Results of FKD on ImageNet ReAL [1] and ImageNetV2 [28] with ResNet-{50, 101}. * indicates that the results are tested using their provided pre-trained model.

Method	ImageNet	ReAL	ImageNetV2 Top-images	ImageNetV2 Matched-frequency	ImageNetV2 Threshold-0.7
<i>ResNet-50:</i>					
ReLabel [52]	78.9	85.0	80.5	67.3	76.0
FKD	79.8	85.5	81.0	68.1	76.7
<i>ResNet-101:</i>					
ReLabel [52]*	80.7	86.5	82.4	69.7	78.2
FKD	81.7	87.0	83.1	70.5	78.9

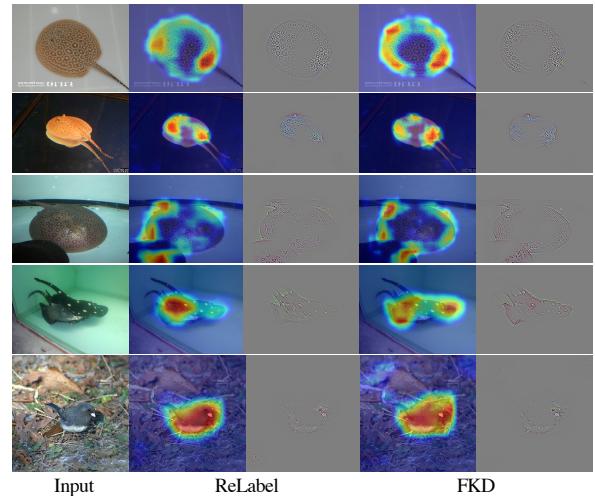


Figure 7. More visualizations of response/attention maps.

C. More Visualizations

We provide more visualizations of the intermediate attention maps from ResNet-50 to explore the learned differences of information between ReLabel and FKD, as shown in Fig. 7. The observation is consistent to our main text.