This CVPR 2020 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Distilling Cross-Task Knowledge via Relationship Matching

Han-Jia Ye
Nanjing University
yehj@lamda.nju.edu.cn

Su Lu
Nanjing University
lus@lamda.nju.edu.cn

De-Chuan Zhan
Nanjing University
zhandc@lamda.nju.edu.cn

Abstract

The discriminative knowledge from a high-capacity deep neural network (a.k.a. the "teacher") could be distilled to facilitate the learning efficacy of a shallow counterpart (a.k.a. the "student"). This paper deals with a general scenario reusing the knowledge from a cross-task teacher — two models are targeting non-overlapping *label spaces. We emphasize that the comparison ability between instances acts as an essential factor threading knowledge across domains, and propose the* ***RE****lationship* ***F****ac****I****litated* ***L****ocal c****L****assifi****E****r* ***D****istillation (*REFILLED*) approach, which decomposes the knowledge distillation flow into branches for embedding and the top-layer classifier. In particular, different from reconciling the instance-label confidence between models,* REFILLED *requires the teacher to* reweight *the hard triplets push forwarded by the student so that the similarity comparison levels between instances are matched. A local embedding-induced classifier from the teacher further supervises the student's classification confidence.* REFILLED *demonstrates its effectiveness when reusing cross-task models, and also achieves state-of-the-art performance on the standard knowledge distillation benchmarks. The code of the paper can be accessed at* https://github.com/njulus/ReFilled.

1. Introduction

Knowledge distillation [6, 20, 64] facilitates the learning efficiency of a deep neural network by reusing the "dark knowledge" from another model. In detail, a strong classifier, e.g., a neural network trained with deeper architectures [43], high-quality images [65], or precise optimization strategies [13, 60], acts as a "teacher", and guides the training of a weaker "student" model. Such model-based knowledge reuse improves the discriminative ability of the target student model, and relieves the burden of model training and storage as well [20, 43, 64, 13]. Its success has been witnessed in a wide range of applications such as model/dataset compression [56, 2, 35, 36, 8], multi-task learning [68, 27], incremental image classification [69, 24].

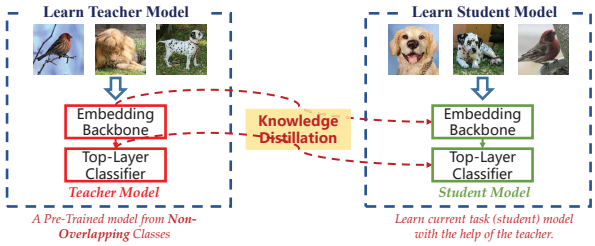

Figure 1. An illustration of reusing the knowledge from a *Cross-Task* teacher model. In a classification task, a teacher is learned from images with non-overlapping classes, while its learning experience is distilled to facilitate the training of the student model.

The main idea of knowledge distillation is to specify a kind of dark knowledge, based on which the student is asked to align with the teacher. For example, the teacher specifies the proportion of how much similar an instance with candidate categories rather than the extreme "black or white" supervision. Although the predictions matching enables the transition of knowledge flow across different neural architectures [20, 35], its dependence on the instance-label relationship restricts both teacher and student to the same label space. In this paper, we enable the student to utilize the learning experience from a *cross-task* teacher, i.e., a neural network with non-overlapping label spaces, which generalizes the knowledge reuse approaches to more applications.

The label difference between teacher and student impedes the direct learning experience transition [22]. The *comparison ability of the embeddings* — measuring how similar two instances are — captures a kind of invariant nature of the model [1] and is free from the label constraint [47, 33, 22]. For a teacher and a student discerning 'Husky vs. Birman" and "Poodle vs. Persian" respectively, the teacher's discriminative embedding encoding the "dog-cat" related characteristics is capable of estimating the similarity relationship of instances in the student's task. Thus, we emphasize the *instance-instance relationship* to bridge the knowledge transfer across different tasks, and thread the knowledge reuse for both embedding and top-layer classifier by taking advantage of the teacher's comparison ability. Figure 1 illustrates the notion of cross-task distillation.

To this end, we propose a 2-stage approach **RE**lationship

Facilitated **L**ocal c**L**assifi**E**r **D**istillation (REFILLED). First, the discriminative ability of features is stressed. For those hard triplets determined by the student's embedding, how teacher comparing them acts as the soft supervision. In other words, the teacher promotes the discriminative ability of the student's embedding by specifying the proportion for each object *how much* a dissimilar impostor should be far away from a target nearest neighbor. Next, the teacher constructs soft supervisions for classifying each instance by measuring its similarity to a local embedding center. Specifically, the classification confidences of the student model and the embedding-induced "instance-label" predictions of the teacher are aligned. Empirical results verify that the REFILLED effectively transfers the classification ability from a cross-task teacher to a student. The same mechanism obtains the state-of-the-art performance on standard knowledge distillation benchmarks as well. We also investigate the middle-shot learning problem, and REFILLED is superior to some popular meta-learning methods.

In summary, We contribute to enhancing the training efficiency of a deep neural network by reusing the knowledge from a cross-task model. The proposed REFILLED approach aligns the high-order comparison relationship between models in a local manner, and works well in both cross-task and same-task distillation problems.

We start by introducing the related literature and the preliminary in Section 2 and Section 3. Then we formalize our REFILLED approach in Section 4. After detailed discussions, finally are experiments and conclusion.

2. Related Work

It is an effective way to take advantage of the learning experience from related pre-trained models to facilitate the model training in the current task [70]. Different from fine-tuning [16] or parameter regularization [28, 11, 30, 48, 62], knowledge distillation/reuse extracts kinds of dark knowledge/privileged information [53, 51, 52] from a fixed strong model (a.k.a. "teacher") and enrich the target model (a.k.a. student) training with more signals. Distilling the knowledge from one model to another has been investigated for model interpretability [71] and compression [6], which is widely applied between deep neural networks since [20, 44, 35] with the help of soft targets. The teacher is usually set as a high-capacity deep neural network or a previous model generation in the current task [5, 13, 60]. Richer supervisions like hidden layer activations [43, 67, 9, 25], parameter flows [64], and transformations [29] are explored. Theoretical analyses and empirical studies of knowledge distillation could be found in [40, 15, 19, 8].

Owing to the strong correspondence between classifier and categories, it is difficult to reuse the classification knowledge from a cross-task teacher. Heterogeneous transfer learning or multi-task learning train a joint model on current and related domains/tasks to fill the gap of label/distribution divergence [27]. Heterogeneous model reuse takes advantage of the model from a related task, which relieves the burden of data storage so as to decrease the risk of privacy leaking [62, 59]. Meta-learning has also been utilized to transfer knowledge across different label spaces, e.g., the few-shot learning [54, 46, 12, 41], but it requires a special training strategy of the teacher.

Different from matching the instance-label predictions between models, embedding [7, 2], pairwise distance [39, 49], and similarity graph [31] have been investigated to improve the quality of the feature towards discriminative embeddings, so that the "downstream" cross-task clustering and representation learning tasks could be improved [22, 38, 65]. The proposed REFILLED approach is general for both same-task and cross-task distillation, where the classification ability of the teacher is transferred to the student by matching the high-order local comparisons.

Embedding learning improves the feature representation by pulling similar instances together and pushing dissimilar ones away [57, 45, 33, 63]. Kinds of side-information such as pairs [10] and triplets [57] are collected as weak supervision in terms of the instance-wise relationship. Stochastic embeddings [32, 50, 3] learn hidden representation to explain the provided relationships, and in REFILLED, the relative instance comparisons measured by a cross-task teacher model is embedded by the student. A local version of the nearest center mean classifier [34, 46] is leveraged to distill the classification ability once with good features.

3. Knowledge Reuse via Distillation

In this section, we first introduce the way to distill knowledge from a high-capacity teacher classifier with soft labels and then describe the cross-task distillation problem.

3.1. Background and Notations

For a C-class classification task, we denote the training data with N examples as $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^D$ and $\mathbf{y}_i \in \{0,1\}^C$ are instance and one-hot label, respectively. Index of 1 in $\mathbf{y}_i$ indicates the class of $\mathbf{x}_i$. The target is to learn a classifier $f(\mathbf{x})\colon \mathbb{R}^D \mapsto \{0,1\}^C$ (e.g., a deep neural network) based on $\mathcal{D}$, which maps an instance to its label. f could be decomposed into a feature extractor $\phi\colon \mathbb{R}^D \mapsto \mathbb{R}^d$ and a linear classifier $W \in \mathbb{R}^{d\times C}$, such that $f(\mathbf{x}) = W^\top \phi(\mathbf{x})$.[1] The objective to learn the model f is

$$\min_f \sum_{i=1}^N \ell\left(f(\mathbf{x}_i),\ \mathbf{y}_i\right) \tag{1}$$

ℓ is the loss such as the cross-entropy, which measures the discrepancy between the prediction and the true label.

[1] We omit the bias term for discussion simplicity.

3.2. Distill Knowledge from a Same-Task Teacher

To improve the training efficacy of f, [20] suggests to distill "dark knowledge" from another pre-trained teacher model via aligning the soft targets:

$$\min_{f_S} \sum_{i=1}^{N} \ell(f_S(x_i), \mathbf{y}_i) + \lambda \mathcal{R}(\mathbf{s}_\tau(f_T(\mathbf{x}_i)), \mathbf{s}_\tau(f_S(\mathbf{x}_i))) \tag{2}$$

Subscripts "T" and "S" denote the model/parameters of the teacher and student (the current task model), respectively. $\lambda > 0$ is a trade-off parameter. $\mathbf{s}_\tau$ transforms the logit into a softened C-way probability:

$$\mathbf{s}_\tau(f(\mathbf{x}_i)) = \mathbf{softmax}(\frac{f(\mathbf{x}_i)}{\tau}) \tag{3}$$

τ is an non-negative temperature, the larger the value of τ, the smoother the output. $\mathcal{R}(\cdot)$ measures the difference between two distributions, e.g., the Kullback-Leibler divergence. In Eq. 2, the student not only minimizes the mapping f from an instance to its label over $\mathcal{D}$, but also keeps its predictions consistent with the teacher. Note that the student and the teacher could use different temperatures.

Since the teacher model usually possesses larger capacity [20, 7, 35] or better parameters [13, 60], its predictions encode the relationship between an instance and its candidate classes. Other forms of dark knowledge along the thread of instance-label mapping are also investigated, such as hidden activation [43] and parameter flows [64].

3.3. Distill Knowledge from a Cross-Task Teacher

The knowledge reuse in Eq. 2 requires the teacher network to target the same labels as the student model so that their classification results on the same instance could be matched. While in a general scenario, it is necessary to borrow the learning experience from a *cross-task* teacher, i.e., a pre-trained teacher f_T on *non-overlapping classes* with the student f_S. The relaxing of the learning condition enables knowledge reuse across related tasks.

4. REFILLED for Cross-Task Distillation

We introduce the main idea of **RE**lationship **F**ac**I**litated **L**ocal c**L**assifi**E**r **D**istillation (REFILLED) approach, followed by analysis and discussions of its two stages.

4.1. Main Ideas of REFILLED

Towards reusing the knowledge from a cross-task teacher, REFILLED decomposes the model into two components, i.e., the embedding and the top-layer classifier, such that the knowledge for each component could be distilled respectively. There are two stages in REFILLED. The discriminative ability of features is distilled through aligning the high-order instance-wise comparisons of the student with the teacher, which bridges the gap between non-overlapping label spaces. After that, the teacher's classification confidences based on local embedding centers further facilitates the classifier training of the student.

4.2. Distill the Embedding

Empirical studies verify the embedding extracted by the penultimate layer of a deep neural network possesses discriminative property [58, 18, 1], where similar instances are close and dissimilar ones are far away. Since instance embedding reveals whether two objects are similar or not, and does not rely on the specific label of each class, thus it could be used across different label spaces [57, 4, 45, 47, 33, 22].

Direct Embedding Distillation. One intuitive way to match the instance-wise relationship between teacher and student is to align their embeddings directly, e.g., minimizing the loss $\|\phi_S(\mathbf{x}) - \phi_T(\mathbf{x})\|_2^2$ over all instances in the current task [7, 14, 25]. This constraint requires both models to have the same size of embeddings, which is too strong to satisfy especially there exists an architecture gap between two models. [31, 38, 39, 49] reuse the embedding-based pairwise relationship of the teacher, where the pairwise similarity measured by the student's embedding should have the same value as the teacher's measure. It still suffers the architecture difference —- even the student has the right similarity relationship, it could still be wrongly rectified by the teacher due to their scale differences. Therefore, considering the discrepancy between the embedding spaces, in REFILLED, we ask the teacher to provide its estimation about *relative comparisons* among instances in the form of *triplets* and require the student to align such relative similarity determination to obtain discriminative embeddings.

Align Triplet. A triplet $(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k)$ contains an anchor $\mathbf{x}_i$, its similar target neighbor $\mathbf{x}_j$, and its dissimilar impostor $\mathbf{x}_k$.[2] The distance between $(\mathbf{x}_i, \mathbf{x}_j)$ based on the embedding ϕ is $\mathrm{Dist}_\phi(\mathbf{x}_i, \mathbf{x}_j) = ||\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)||_2$. A good embedding makes $\mathrm{Dist}_\phi(\mathbf{x}_i, \mathbf{x}_j)$ smaller than $\mathrm{Dist}_\phi(\mathbf{x}_i, \mathbf{x}_k)$. We use the stochastic triplet probability [50] as a kind of "dark knowledge", which encodes how much the anchor is close to its target neighbor than its impostor:

$$p_{ijk}(\phi) = \frac{\exp\left(-\mathrm{Dist}_\phi\left(\mathbf{x}_i, \mathbf{x}_j\right)/\tau\right)}{\exp\left(-\mathrm{Dist}_\phi\left(\mathbf{x}_i, \mathbf{x}_j\right)/\tau\right) + \exp\left(-\mathrm{Dist}_\phi\left(\mathbf{x}_i, \mathbf{x}_k\right)/\tau\right)} \tag{4}$$

Eq. 4 measures the relative instance-wise similarities in a triplet form. If the target neighbor $\mathbf{x}_j$ is close to the anchor while the impostor is far away, p_{ijk} is large, otherwise

[2] Usually, we think two instances are similar if they come from the same class, and they are dissimilar if they have different labels.

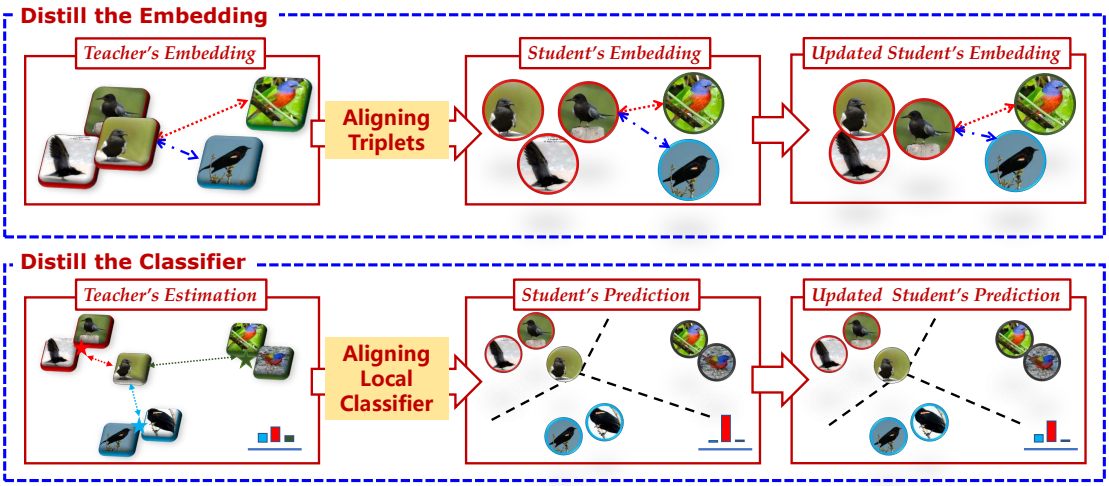

Figure 2. Illustration of the proposed **RE**lationship **F**ac**I**litated **L**ocal c**L**assifi**E**r **D**istillation (REFILLED), which has two stages: it first distills the discriminative embedding by aligning triplets, e.g. the relative similarities between two impostors (denoted by the red and blue arrow) are specified by the teacher; REFILLED then distills the classification ability via local embedding-based classifiers. With the class prototype (denoted by stars), the teacher provides a good estimation for the classification confidence. More details can be found in the text.

the probability is small. Different from the vanilla triplets generated from labels with only the "similar or not" binary information [57, 45, 47, 33], we take advantage of the stochastic triplet probability to introduce richer similarity comparison information towards more effective embedding learning. With a bit abuse of the notation, we also use the temperature τ to soften the probability in Eq. 4.

In REFILLED, we improve the discriminative ability of the student model embedding ϕ_S by distilling the triplet comparison knowledge from the teacher. Define the Bernoulli distribution $\mathcal{P}_{ijk}(\phi) = [p_{ijk}(\phi), 1 - p_{ijk}(\phi)]$, we minimize the KL-divergence over all generated triplets:

$$\min_{\phi_S} \sum_{ijk} \mathbf{KL}\big(\mathcal{P}_{ijk}(\phi_T) \parallel \mathcal{P}_{ijk}(\phi_S)\big) \tag{5}$$

By aligning the novel kind of dark knowledge in Eq. 5, the student is expected to have better comparison ability as strong as the teacher. There are two main advantages of the triplet matching. With the help of the teacher, Eq. 5 not only encodes the high-order relationship between instances but also specifies the differences between the generated triplets. For example, although three images of "black tern" are similar to one "red-winged black bird" image, the two flying black terns should be more close than the one black terns drinking the water. Besides, aligning the triplet comparisons between different models gets rid of the scale and embedding size differences between neural architectures.

It still remains one key component of collecting the triplets for relationship distillation. In our implementation, we generate "semi-hard" triplets [45] based on the student's embedding (the triplets with relatively smaller $\text{Dist}_{\phi_S}(\mathbf{x}_i, \mathbf{x}_k)$ than $\text{Dist}_{\phi_S}(\mathbf{x}_i, \mathbf{x}_j)$). Thus, if the student finds some triplets hard to evaluate, it will query the teacher for concrete measures of the similarity proportions. We do ℓ_2-normalization on all the embeddings before computing their distances, and only apply the temperature in $\mathcal{P}_{ijk}(\phi_T)$.

Discussions. Define $\rho_{ijk} = 1 - p_{ijk}(\phi_T)$ and $\iota(x) = \ln(1 + \exp(-x))$ as the logistic loss, we can rethink the objective in Eq. 5 by reformulating

$$\begin{aligned} & \mathbf{KL}\left(\mathcal{P}_{ijk}(\phi_T) \parallel \mathcal{P}_{ijk}(\phi_S)\right) \\ \cong\ & \rho_{ijk}\left(\text{Dist}_{\phi_S}(\mathbf{x}_i, \mathbf{x}_k) - \text{Dist}_{\phi_S}(\mathbf{x}_i, \mathbf{x}_j)\right) \\ & + \iota\left(\text{Dist}_{\phi_S}(\mathbf{x}_i, \mathbf{x}_k) - \text{Dist}_{\phi_S}(\mathbf{x}_i, \mathbf{x}_j)\right) \end{aligned} \tag{6}$$

$\cong$ neglects the constants. In addition to optimizing the embedding triplets with the loss ι, Eq. 6 *adds different weights* when minimizing (resp. maximizing) the distance between similar (resp. dissimilar) pairs based on the teacher's estimation. For example, if $(\mathbf{x}_i, \mathbf{x}_j)$ are not too similar compared with $(\mathbf{x}_i, \mathbf{x}_k)$, the teacher will specify a relative lower probability p_{ijk} to compensate for the over-emphasizing of similarity/dissimilarity in the triplet, and the minimization of $\text{Dist}_{\phi_S}(\mathbf{x}_i, \mathbf{x}_j)$ in ι is weakened with weight ρ_{ijk} .

4.3. Distill the Local Classifier

The student's embedding depicts the relationship between instances as well as the teacher by aligning the triplet probabilities, which facilitates the "downstream" task. Considering the transition between label space, REFILLED further proposes to distill the classification ability from the teacher via an embedding-based local classifier.

Embedding-Based Local Classifier. During the stochastic optimization of the student model, with a little abuse of notations, denote $(X \in \mathbb{R}^{N\times D}, Y \in \{0,1\}^{N\times C})$ as the

instances and one-hot labels in the sampled mini-batch, respectively. Note that the batch may not cover all the classes in the data. With the teacher's embeddings $\phi_T(X) \in \mathbb{R}^{N\times d}$ on X, we compute the embedding center of each class by

$$P = \mathrm{diag}(\mathbf{1} \oslash (Y^\top \mathbf{1}))Y^\top \phi_T(X) \in \mathbb{R}^{C\times d} \tag{7}$$

$\oslash$ denotes the element-wise division. Each row $\mathbf{p}_c \in \mathbb{R}^d$ of P corresponds to the center of the c-th class in the batch. The label of an instance in the batch can be determined by:

$$p_{\phi_T}(\mathbf{y}_i \mid \mathbf{x}_i) = \mathbf{softmax}\left(-\|\phi_T(\mathbf{x}_i) - \mathbf{p}_c\|_2^2/\tau\right) \tag{8}$$

which is normalized over the set of negative distances from the instance embedding $\phi(\mathbf{x}_i)$ to all class centers in P. $p_{\phi_T}(\mathbf{y}_i \mid \mathbf{x}_i)$ is large if $\phi_T(\mathbf{x}_i)$ is close to $\mathbf{p}_c$ in the teacher's embedding space. τ is the temperature. Eq. 8 works in the same manner as a local nearest center mean classifier [34], where only the classes in the current sampled batch are taken into account. It can be applied even to the classification tasks across non-overlapping label spaces [46, 61].

Local Knowledge Distillation. Equipped with Eq. 8, the classification ability of a cross-task teacher could be further reused for training the student's classifier. Therefore, we incorporate a local knowledge distillation term with Eq. 1:

$$\min_{f_S} \sum_{i=1}^{N} \ell\left(f_S(x_i), \mathbf{y}_i\right) + \lambda \mathbf{KL}\big(p_{\phi_T}(\mathbf{y}_i \mid \mathbf{x}_i),\ \mathbf{s}_\tau(f_S(\mathbf{x}_i))\big) \tag{9}$$

Benefited from the local classifier induced from the teacher's embedding, the classifier of the student could be further supervised by a cross-task teacher. In the second term of Eq. 9, rather than aligning two model's confidences of all classes in the data set, only the posteriors of classes in the sampled mini-batch are matched. This local knowledge helps when distilling from a same-class teacher as well (refer to Section 5.2), where two models match predictions over the sampled classes in the mini-batch. In the implementation, we also investigate an exponential-decayed weight to set λ, so that the student relies on the teacher's supervision during its initial learning period while weakening the teacher's guide if itself is strong enough.

Discussions. By decoupling the embedding ϕ and the linear classifier W, the effectiveness of the knowledge distillation could be analyzed by its gradient over the classifier $\mathbf{w}_c$ of the c-th class (denote the objective of Eq. 2 as O):

$$\frac{\partial O}{\partial \mathbf{w}_c} = \sum_{\mathbf{x}} \left[-p_c + \sum_{c'=1}^{C} p_{c'} q_c\right] \phi(\mathbf{x}) \tag{10}$$

q_c and p_c are the teacher's and student's posterior probabilities of the c-th class given instance $\mathbf{x}$, respectively. Different from the vanilla loss, when considering the soft supervision from the teacher, not only the instance from the target class but also those from helpful related classes (the ones with large $p_{c'}$) will be incorporated to direct the update of the classifier. Since the summation in Eq. 10 is computed over all C classes, the normalized class posterior q_c becomes small if C is large, so that the helpful class instance will not be stressed obviously. Therefore, we consider a *local* version of the knowledge distillation term in Eq. 9, where only the classes in the current mini-batch are considered, i.e., the influence of a helpful related class selected by the teacher will be better emphasized in the update of $\mathbf{w}_c$.

Algorithm 1 The Flow of REFILLED.

Require: Pre-trained Teacher's Embedding ϕ_T.
Distill the Embedding:
for all Iter = 1,...,MaxIter **do**
  Sample a mini-batch $\{(\mathbf{x}_i, \mathbf{y}_i)\}$.
  Generate triplets $\{(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k)\}$ with student's embeddings $\{\phi_S(\mathbf{x})\}$.
  Compute probability of triplets $p_{ijk}(\phi_T)$ as Eq. 4.
  Optimizing ϕ_S by aligning triplets in Eq. 5.
end for
Distill the Classifier:
Initialize f_S with ϕ_S.
Optimizing f_S with Eq. 9.

The Two-Stage REFILLED Approach. In summary, there are two steps in REFILLED to reuse the holistic knowledge of the teacher through its embedding, so that to improve the discerning ability of the student's embedding and classifier, respectively. The whole flow of REFILLED for cross-task distillation is illustrated in Figure 2 and Alg. 1.

5. Experiments

We verify REFILLED on a variety of tasks, namely classification by reusing cross-task models, standard knowledge distillation, and middle-shot learning.

5.1. Cross-Task Knowledge Distillation

REFILLED is able to reuse a cross-task teacher to assist the training of a student model.

Datasets. Caltech-UCSD Birds-200-2011 (CUB) [55] constructs a fine-grained classification problem over 200 different species of birds. We use the first 100 classes to train the teacher, and learn the student model on the remaining 100 classes.

Implementation Details. We use different configurations of the MobileNets [21] and adjust the model complexity

Table 1. The mean accuracy of cross-task distillation on CUB data set, where teacher and student are trained for non-overlapping 100 classes with MobileNets. The three values in the "teacher" row correspond to baselines: applying 1NN based on teacher's embedding, train a linear LR classifier based on fixed teacher's embedding, and Fine-Tune (FT) based on teacher's embedding.

Width Multiplier	1	0.75	0.5	0.25
Teacher	1NN: 45.31 , LR: 53.82 , FT: 65.72			
Student	71.25	67.56	66.85	64.48
RKD [38]	70.83	68.80	67.44	63.97
REFILLED	**73.38**	**70.42**	**69.77**	**67.10**

with different channels (complicated models have larger channels). The teacher is trained with cross-entropy loss and width multiplier 1.0. We change the width multiplier of the student in $\{1, 0.75, 0.5, 0.25\}$.

Evaluations. For each 100-way classification task, we split 70% of data in each class for training, and the remaining is used for test. The teacher model is first trained on the first 100 classes till convergence and then used to direct the training of the student model upon *non-overlapping* classes. The averaged classification accuracy over 3 trials is reported. The neural networks are optimized by SGD w/ momentum. Detailed configurations are in the supp.

Results. The results of cross-task distillation are in Table 1. We first investigate three baselines by adapting the teacher for cross-task classification, i.e., the 1NN based on teacher's embedding, training a linear Logistic Regression (LR) upon the fixed teacher's embedding, and Fine-Tuning (FT) the teacher model initialized by the pre-trained embedding. The test accuracy of the student becomes higher when learning the task with more complicated models (w/ larger width multiplier value). We also compare with one representative embedding-based approach Relation Knowledge Distillation (RKD) [38], and fine-tune the model after obtaining the distilled embedding from the cross-task teacher. RKD sometimes gets better accuracy than the vanilla student model. Our REFILLED achieves the best classification performance in all cases. Benefited from reusing the knowledge from the teacher, the classification achieves a further improvement w.r.t. the vanilla training.

Will All Components in REFILLED Help? Given pre-trained weights of the teacher and fixing the width multiple of the student equals 1, we investigate three fine-tuning variants in Figure 3 besides training the student model directly (Vanilla) , namely, fine-tuning with the distilled embedding after the first stage of REFILLED (REFILLED^{1st}), fine-tuning with Eq. 9 using fixed λ (REFILLED^{-}), and REFILLED (which has exponential-decayed λ). The step-wise improvements of the classification results verify the effectiveness of each component in REFILLED.

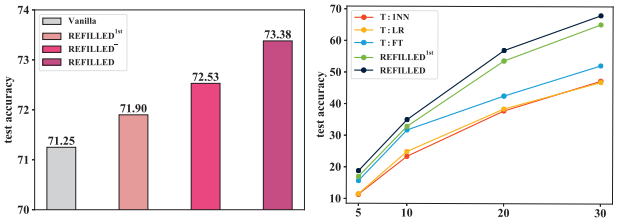

Figure 3. Left: The mean accuracy of different variants of REFILLED on CUB for cross-task distillation; Right: The change of accuracy when the number of instances per class (shot) varies.

REFILLED with Different Size of Target Task Data. To test the extreme of the knowledge distillation ability of REFILLED, we construct the target classification task with different sizes of training data. When the number of effective training data is small, it is more difficult to train the student model, so that the help from the teacher becomes more important. We vary the number of instances per class in the student's task from 5 to 30, and the averaged classification accuracies are shown in Figure 3. REFILLED keeps a performance margin with comparison methods in all cases.

5.2. Standard Knowledge Distillation

The REFILLED is a general approach that helps the training of a student with a same-class teacher.

Datasets. Following [2], we test the knowledge distillation ability of REFILLED on another benchmark CIFAR-100 [26] besides CUB. CIFAR-100 contains 100 classes with 6000 small images per class. In each class, there are 5,000 images for training and 1,000 images for test. We use the standard split to train both teacher and student models. We also evaluate REFILLED on CUB, where all 200 classes are used during training based on the standard split.

Implementation Details. We test the effectiveness of REFILLED across diverse architectures, i.e., ResNet [17], Wide-ResNet [66], and MobileNets [21]. Towards investigating different capacities of the teacher and student, we change the depth of ResNet (through the number of layers), the width and depth of Wide-ResNet, and the width multiplier of MobileNets. Both teacher and student are trained on the same training set till convergence.

Evaluations. Both teacher and student are trained on the same set with three different seeds of initialization, and we report the mean accuracy of the student on the test set.

Table 2. The average classification results of knowledge distillation methods on CIFAR-100 data set based on the Wide-ResNet. We fix the teacher with (depth, width) $= (40, 2)$, and set the student capacity with different (depth, width) values.

(depth, width)	(40, 2)	(16, 2)	(40, 1)	(16, 1)
Teacher	74.44			
Student	74.44	70.15	68.97	65.44
KD [20]	75.47	71.87	70.46	66.54
FitNet [43]	74.29	70.89	68.66	65.38
AT [67]	74.76	71.06	69.85	65.31
NST [23]	74.81	71.19	68.00	64.95
VID-I [2]	75.25	73.31	71.51	66.32
KD+VID-I [2]	76.11	73.69	72.16	67.19
RKD [38]	76.62	72.56	72.18	65.22
REFILLED	**77.49**	**74.01**	**72.72**	**67.56**

Table 3. The average classification results of knowledge distillation methods on CUB based on MobileNets. We fix the teacher's width multiplier to 1.0, and change the student's multipliers.

Width Multiplier	1	0.75	0.5	0.25
Teacher	75.36			
Student	75.36	74.87	72.41	69.72
KD [20]	77.61	76.02	74.24	72.03
FitNet [43]	75.10	75.03	72.17	69.09
AT [67]	76.22	76.10	73.70	70.74
NST [23]	76.91	77.05	74.03	71.54
KD+VID-I [2]	77.03	76.91	75.62	72.23
RKD [38]	77.72	76.80	74.99	72.55
REFILLED	**78.95**	**78.01**	**76.11**	**73.42**

Distillation From Same Architecture Family Models. We first test the case when teacher and student come from the same model family. The results on CIFAR-100 and CUB could be found in Table 2 and Table 3, respectively. On CIFAR-100 we exactly follow the evaluation protocol in [2], which implements teacher and student with the Wide-ResNet. We re-implement RKD [38] and cite the results of other comparison methods from [2]. For CUB, we use MobileNets as the basic model. Since the teacher possesses more capacity, its learning experience assists the training of the student once utilizing the knowledge distillation methods. REFILLED achieves the best classification performance in all settings, which validates transferring the knowledge for both embedding and classifier is one of the key factors for model reuse.

Will Embedding Help for Knowledge Distillation? We use the Normalized Mutual Information (NMI) as a criterion to measure the embedding quality, the larger the better. In Table 4, we compute NMI for student model's embedding trained with and without aligning the teacher's triplets in CIFAR-100. Figure 4 visualizes the embedding quality over 10 sampled classes using tSNE [50]. Both quantitative and qualitative results verify the effectiveness of the triplet aligning step in REFILLED for knowledge distillation.

Table 4. The NMI on CIFAR-100 to evaluate the embedding quality before and after the Triplet Aligning (TA) step in REFILLED.

(depth, width)	(40, 2)	(16, 2)	(40, 1)	(16, 1)
w/o TA	56.50	54.91	54.02	51.77
w/ TA	59.63	57.98	57.62	54.39

Table 5. The mean accuracy on CIFAR-100 to evaluate the effectiveness of Local Knowledge Distillation (LKD) in REFILLED.

(depth, width)	(40, 2)	(16, 2)	(40, 1)	(16, 1)
w/ KD	77.08	73.57	72.24	67.14
w/ Local KD	77.49	74.01	72.72	67.56

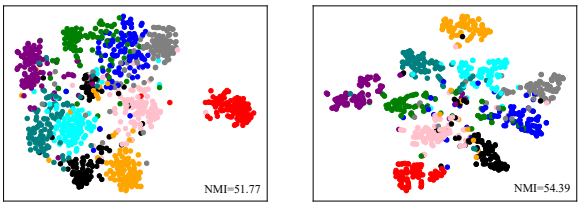

Figure 4. The tSNE [50] of the vanilla student training (left) and the improved embedding after the 1st stage of REFILLED (right) over 10 classes sampled from CIFAR-100.

Will Local Knowledge Distillation Help? Results in Table 5 verify the further improvement of Local Knowledge Distillation (LKD) in Eq. 9 compared with the vanilla Knowledge Distillation (KD) when training based on the distilled embedding after the first stage of REFILLED. A local consideration of probability matching helps.

Distillation From Different Model Families. To further evaluate the performance of REFILLED, we use REFILLED to distill the knowledge from a cross-family teacher. For CIFAR-100, we set the teacher as ResNet-110, and use the MobileNets with different channels as the student model. Table 6 demonstrates the results, and REFILLED keeps its superiority in this case. More results are in the supp.

5.3. Middle-Shot Learning

Training a deep neural network with middle-shot of data is a difficult task, where models are prone to over-fit. In this subsection, we apply our REFILLED approach for middle-shot learning, where the classification ability from a teacher trained on SEEN class can be used to help the student model training for UNSEEN middle-shot tasks.

Table 6. The average classification accuracy of knowledge distillation methods on CIFAR-100 data set. The teacher is trained with ResNet-110, which gets 74.09% test accuracy. The student is learned with MobileNets, whose width multiplier is changed.

Width Multiplier	1	0.75	0.5	0.25
Student	68.57	67.92	65.66	60.87
KD [20]	70.34	68.21	66.06	61.38
FitNet [43]	67.99	67.85	65.12	61.01
AT [67]	68.97	67.88	66.44	62.15
NST [23]	70.62	70.49	69.15	61.32
KD+VID-I [2]	71.94	70.13	68.51	62.50
RKD [38]	70.41	68.93	66.24	61.44
REFILLED	**73.81**	**72.88**	**70.02**	**63.15**

Datasets. We use the popular *Mini*ImageNet data set [54], which contains 100 classes and 600 images in each class. Following [54, 42], there are 64 classes (SEEN class) to train the teacher, 16 classes for validation, and we sample tasks from the remaining 20 classes to train the student.

Implementation Details. Following the literature, we investigate two different backbones, a 4-layer ConvNet [54, 46, 12] and the ResNet [37, 61], which outputs embeddings with 64 and 640 dimensions, respectively. We train a teacher model on the SEEN classes with ResNet or ConvNet, and use the teacher model to help the training of the student classifier on tasks composed by UNSEEN classes.

Evaluations. Define a K-shot C-way task as a C-class classification problem with K instances per class. Different from the few-shot learning setting where $K \in \{1, 5\}$, here we consider there are a bit more instances in each class, i.e., $K = \{10, 30\}$. Note that even $K = 30$ is not enough to train a complicated neural network from scratch. We sample 5-way tasks from the 20-class split to train the student model and evaluate its performance by classifying another 15 instances from each of the 5 sampled classes. We evaluate the final performance by mean accuracy over 600 trials. More results of few-shot learning are in supp.

Comparison Methods. Meta-learning is a popular way to solve the few-shot classification problem. To mimic the test case, it samples C-Way K-Shot tasks from the SEEN class set to learn task-level inductive bias like embedding [54, 46] or initialization [12, 41]. However, the computational burden (e.g., the batch size) becomes large when the number of shots increases. Besides, meta-learning needs to specify the way to obtain a meta-model from the SEEN classes. We compare our methods with the embedding-based meta-learning approaches like and ProtoNet [46] and FEAT [61]. We can make predictions directly with the teacher's embedding, the penultimate layer of the teacher, by leveraging the nearest neighbor (1NN). Based on the teacher's embedding, we also train linear classifiers like SVM or fine-tune the whole model upon the middle-shot training data of sample tasks. We tune the hyper-parameters of such methods with sampled middle-shot tasks from the validation split.

Table 7. The mean accuracy over 600 trials of middle-shot tasks. We set the student model as the ConvNet, and investigate both ResNet and ConvNet as the teacher model, for our REFILLED approach. Detailed results and configurations are in the supp. REFILLED^1 denotes the result reusing a ResNet teacher and REFILLED^2 stands for the result reusing a ConvNet teacher.

Tasks	10-Shot 5-Way	30-Shot 5-Way
1NN	66.56	69.80
SVM	74.24	77.87
Fine-Tune	74.95	78.62
ProtoNet [46]	74.42	78.10
FEAT [61]	74.86	78.84
REFILLED^1	**76.42**	**80.33**
REFILLED^2	**75.37**	**78.94**

Results. The results of middle-shot learning are shown in Table 7. When the number of shots becomes large, fine-tuning is a very strong baseline, which gets better results than some meta-learning approaches. Our REFILLED method achieves better results than fine-tune, which validates the importance of reusing the knowledge of a cross-task teacher for training a classifier.

6. Conclusion

Although knowledge distillation facilitates the transition of learning experience between heterogeneous models, i.e., neural networks with different architectures, it is still challenging to reuse models across non-overlapping label spaces. In this paper, we focus on matching the comparison ability on account of embeddings, which not only gets rid of the label space constraint but also captures the high order relationships among instances. The proposed **RE**lationship **F**ac**I**litated **L**ocal c**L**assifi**E**r **D**istillation (REFILLED) approach has two stages, namely embedding aligning and local knowledge distillation. Besides improving the learning efficiency by reusing cross-task models, REFILLED also achieves better classification performance in standard knowledge distillation tasks.

Acknowledgments

This work is partially supported by The National Key R&D Program of China (2018YFB1004300), NSFC (61773198, 61773198, 61632004), and NSFC-NRF joint research project (61861146001).

References

[1] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018. 1, 3

[2] Sungsoo Ahn, Shell Xu Hu, Andreas C. Damianou, Neil D. Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *CVPR*, pages 9163–9171, 2019. 1, 2, 6, 7, 8

[3] Ehsan Amid and Antti Ukkonen. Multiview triplet embedding: Learning attributes in multiple maps. In *ICML*, pages 1472–1480, 2015. 2

[4] Ehsan Amid and Antti Ukkonen. Multiview triplet embedding: Learning attributes in multiple maps. In *ICML*, pages 1472–1480, 2015. 3

[5] Hessam Bagherinezhad, Maxwell Horton, Mohammad Rastegari, and Ali Farhadi. Label refinery: Improving imagenet classification through label progression. *CoRR*, abs/1805.02641, 2018. 2

[6] Cristian Bucila, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *KDD*, pages 535–541, 2006. 1, 2

[7] Hanting Chen, Yunhe Wang, Chang Xu, Chao Xu, and Dacheng Tao. Learning student networks via feature embedding. *CoRR*, abs/1812.06597, 2018. 2, 3

[8] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *ICCV*, pages 4794–4802, 2019. 1, 2

[9] Wojciech M. Czarnecki, Simon Osindero, Max Jaderberg, Grzegorz Swirszcz, and Razvan Pascanu. Sobolev training for neural networks. In *NeurIPS*, pages 4281–4290. 2017. 2

[10] Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S. Dhillon. Information-theoretic metric learning. In *ICML*, pages 209–216, 2007. 2

[11] Simon S. Du, Jayanth Koushik, Aarti Singh, and Barnabás Póczos. Hypothesis transfer learning via transformation functions. In *NeurIPS*, pages 574–584. 2017. 2

[12] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pages 1126–1135, 2017. 2, 8

[13] Tommaso Furlanello, Zachary Chase Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born-again neural networks. In *ICML*, pages 1602–1611, 2018. 1, 2, 3

[14] Mengya Gao, Yujun Shen, Quanquan Li, Chen Change Loy, and Xiaoou Tang. Feature matters: A stage-by-stage approach for knowledge transfer. *CoRR*, abs/1812.01819, 2018. 3

[15] Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation. In *ICLR*, 2019. 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, pages 1026–1034, 2015. 2

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[18] Xinwei He, Yang Zhou, Zhichao Zhou, Song Bai, and Xiang Bai. Triplet-center loss for multi-view 3d object retrieval. In *CVPR*, pages 1945–1954, 2018. 3

[19] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *ICCV*, pages 1921–1930, 2019. 2

[20] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. 1, 2, 3, 7, 8

[21] 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. *CoRR*, abs/1704.04861, 2017. 5, 6

[22] Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. In *ICLR*, 2018. 1, 2, 3

[23] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *CoRR*, abs/1707.01219, 2017. 7, 8

[24] Khurram Javed and Faisal Shafait. Revisiting distillation and incremental classifier learning. In *ACCV*, pages 3–17, 2018. 1

[25] Animesh Koratana, Daniel Kang, Peter Bailis, and Matei Zaharia. LIT: learned intermediate representation training for model compression. In *ICML*, pages 3509–3518, 2019. 2, 3

[26] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 6

[27] Jogendra Nath Kundu, Nishank Lakkakula, and R. Venkatesh Babu. Um-adapt: Unsupervised multi-task adaptation using adversarial cross-task distillation. In *ICCV*, pages 1436–1445, 2019. 1, 2

[28] Ilja Kuzborskij and Francesco Orabona. Fast rates by transferring from auxiliary hypotheses. *Machine Learning*, 106(2):171–195, 2017. 2

[29] Seung Hyun Lee, Dae Ha Kim, and Byung Cheol Song. Self-supervised knowledge distillation using singular value decomposition. In *ECCV*, pages 339–354, 2018. 2

[30] Xuhong Li, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *ICML*, pages 2830–2839, 2018. 2

[31] Yufan Liu, Jiajiong Cao, Bing Li, Chunfeng Yuan, Weiming Hu, Yangxi Li, and Yunqiang Duan. Knowledge distillation via instance relationship graph. In *CVPR*, pages 7096–7104, 2019. 2, 3

[32] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(Nov):2579–2605, 2008. 2

[33] R. Manmatha, Chao-Yuan Wu, Alexander J. Smola, and Philipp Krähenbühl. Sampling matters in deep embedding learning. In *CVPR*, pages 2859–2867, 2017. 1, 2, 3, 4

[34] Thomas Mensink, Jakob J. Verbeek, Florent Perronnin, and Gabriela Csurka. Distance-based image classification: Generalizing to new classes at near-zero cost. *TPAMI*, 35(11):2624–2637, 2013. 2, 5

[35] Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant: Bridging the gap between student and teacher. *CoRR*, abs/1902.03393, 2019. 1, 2, 3

[36] Gaurav Kumar Nayak, Konda Reddy Mopuri, Vaisakh Shaj, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. Zero-shot knowledge distillation in deep networks. In *ICML*, 2019. 1

[37] Boris N. Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. TADAM: task dependent adaptive metric for improved few-shot learning. In *NeurIPS*, pages 719–729. 2018. 8

[38] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, pages 3967–3976, 2019. 2, 3, 6, 7, 8

[39] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *ICCV*, pages 5007–5016, 2019. 2, 3

[40] Mary Phuong and Christoph Lampert. Towards understanding knowledge distillation. In *ICML*, pages 5142–5151, 2019. 2

[41] Hang Qi, Matthew Brown, and David G. Lowe. Low-shot learning with imprinted weights. In *CVPR*, pages 5822–5830, 2018. 2, 8

[42] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017. 8

[43] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *ICLR*, 2015. 1, 2, 3, 7, 8

[44] Bharat Bhusan Sau and Vineeth N. Balasubramanian. Deep model compression: Distilling knowledge from noisy teachers. *CoRR*, abs/1610.09650, 2016. 2

[45] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 2, 3, 4

[46] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, pages 4080–4090. 2017. 2, 5, 8

[47] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *CVPR*, pages 4004–4012, 2016. 1, 3, 4

[48] Suraj Srinivas and François Fleuret. Knowledge transfer with jacobian matching. In *ICML*, pages 4730–4738, 2018. 2

[49] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *ICCV*, pages 1365–1374, 2019. 2, 3

[50] Laurens van der Maaten and Kilian Q. Weinberger. Stochastic triplet embedding. In *MLSP*, pages 1–6, 2012. 2, 3, 7

[51] Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: similarity control and knowledge transfer. *JMLR*, 16:2023–2049, 2015. 2

[52] Vladimir Vapnik and Rauf Izmailov. Learning with intelligent teacher. In *COPA*, pages 3–19, 2016. 2

[53] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5-6):544–557, 2009. 2

[54] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *NeurIPS*, pages 3630–3638. 2016. 2, 8

[55] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 5

[56] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A. Efros. Dataset distillation. *CoRR*, abs/1811.10959, 2018. 1

[57] Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 10:207–244, 2009. 2, 3, 4

[58] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, pages 499–515, 2016. 3

[59] Xi-Zhu Wu, Song Liu, and Zhi-Hua Zhou. Heterogeneous model reuse via optimizing multiparty multiclass margin. In *ICML*, pages 6840–6849, 2019. 2

[60] Chenglin Yang, Lingxi Xie, Chi Su, and Alan L. Yuille. Snapshot distillation: Teacher-student optimization in one generation. In *CVPR*, pages 2859–2868, 2019. 1, 2, 3

[61] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Learning embedding adaptation for few-shot learning. *CoRR*, 2018. 5, 8

[62] Han-Jia Ye, De-Chuan Zhan, Yuan Jiang, and Zhi-Hua Zhou. Rectify heterogeneous models with semantic mapping. In *ICML*, pages 1904–1913, 2018. 2

[63] Han-Jia Ye, De-Chuan Zhan, Yuan Jiang, and Zhi-Hua Zhou. What makes objects similar: A unified multi-metric learning approach. *TPAMI*, 41(5):1257–1270, 2019. 2

[64] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *CVPR*, pages 7130–7138, 2017. 1, 2, 3

[65] Lu Yu, Vacit Oguz Yazici, Xialei Liu, Joost van de Weijer, Yongmei Cheng, and Arnau Ramisa. Learning metrics from teachers: Compact networks for image embedding. In *CVPR*, pages 2907–2916, 2019. 1, 2

[66] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016. 6

[67] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017. 2, 7, 8

[68] Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. Deep mutual learning. In *CVPR*, pages 4320–4328, 2018. 1

[69] Peng Zhou, Long Mai, Jianming Zhang, Ning Xu, Zuxuan Wu, and Larry S. Davis. M2KD: multi-model and multi-level knowledge distillation for incremental learning. *CoRR*, abs/1904.01769, 2019. 1

[70] Zhi-Hua Zhou. Learnware: on the future of machine learning. *FCS*, 10(4):589–590, 2016. 2

[71] Zhi-Hua Zhou and Yuan Jiang. Nec4.5: Neural ensemble based C4.5. *TKDE*, 16(6):770–773, 2004. 2