Distilling from Similar Tasks for Transfer Learning on a Budget

Kenneth Borup
Aarhus University
kennethborup@math.au.dk

Cheng Perng Phoo
Cornell University
cpphoo@cs.cornell.edu

Bharath Hariharan
Cornell University
bharathh@cs.cornell.edu

Abstract

We address the challenge of getting efficient yet accurate recognition systems with limited labels. While recognition models improve with model size and amount of data, many specialized applications of computer vision have severe resource constraints both during training and inference. Transfer learning is an effective solution for training with few labels, however often at the expense of a computationally costly fine-tuning of large base models. We propose to mitigate this unpleasant trade-off between compute and accuracy via semi-supervised cross-domain distillation from a set of diverse source models. Initially, we show how to use task similarity metrics to select a single suitable source model to distill from, and that a good selection process is imperative for good downstream performance of a target model. We dub this approach DistillNearest*. Though effective,* DistillNearest *assumes a single source model matches the target task, which is not always the case. To alleviate this, we propose a weighted multi-source distillation method to distill multiple source models trained on different domains weighted by their relevance for the target task into a single efficient model (named* DistillWeighted*). Our methods need no access to source data and merely need features and pseudo-labels of the source models. When the goal is accurate recognition under computational constraints, both* DistillNearest *and* DistillWeighted *approaches outperform both transfer learning from strong ImageNet initializations as well as state-of-the-art semi-supervised techniques such as FixMatch. Averaged over 8 diverse target tasks our multi-source method outperforms the baselines by 5.6%-points and 4.5%-points, respectively.* ***Code:*** *github.com/Kennethborup/DistillWeighted*

1. Introduction

Recognition models get more accurate the larger they are and the more data they are trained on [21, 36, 45]. This is a problem for many applications of interest in medicine (e.g. X-ray analysis) or science (e.g. satellite-image analysis) where both labeled training data, as well as computational resources needed to train such large models, are lacking.

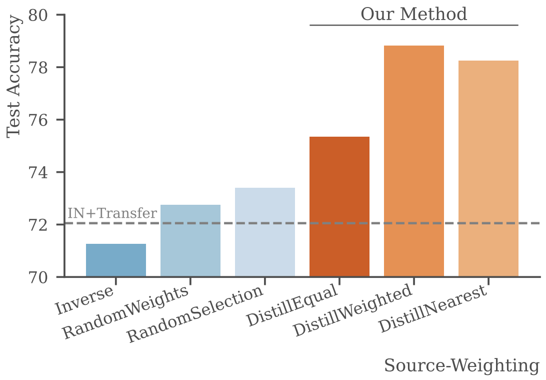

Figure 1: Average test accuracy over five target tasks with different methods for weighting source models for distillation. Our methods outperform the baselines and transfer learning from ImageNet. See Section 5.3 for details.

The challenge of limited labeled data can potentially be alleviated by fine-tuning large-scale “foundation models” [13, 21, 45]. However, fine-tuning is computationally expensive, especially when one looks at foundation models with billions of parameters [13]. Unfortunately, all evidence suggests that larger foundation models perform better at fine-tuning [21, 45]. This leaves downstream applications the unpleasant trade-off of expensive computational hardware for fine-tuning large models, or inaccurate results from smaller models. Motivated by this challenge, we ask *can we train accurate models on tight data and compute budgets without fine-tuning large foundation models?*

To set the scene, we assume the existence of a diverse set (both in architecture and task) of pre-trained source models (or foundation models). We do not have the resources to fine-tune these models, but we assume we can perform inference on these models and extract features, *e.g.* through APIs on cloud services [8, 34]. For the target task, we assume that labeled data is very limited, but unlabeled data is available. We then propose a simple and effective strategy for building an accurate model for the target task: DistillNearest.

Concretely, we first compute a measure of "task similarity" between our target task and each source model and rank the source models accordingly. Then we pseudo-label the unlabeled data using the most similar source model. These pseudo-labels may not even be in the same label space as the target task, but we conjecture that due to the similarity between the source and target tasks, the pseudo-labels will still *group* the target data points in a task-relevant manner. Finally, we train the target model using the pseudo-labels and the available ground truth labeled data. This allows us to bypass the large computations required to fine-tune source models and directly work on the target model. At the same time, we get to effectively use the knowledge of the large source model even if it is trained on a different task.

DISTILLNEAREST assumes that a *single* best source model exists. But for some target tasks, we might need to combine multiple source models to achieve a sufficiently diverse representation to distill. We, therefore, propose an extension of our approach that distills *multiple (diverse) source models* trained on different domains, weighted by their relevance for the target task. This extension obtains even further improvements on our target performance (see Figure 1). We dub this method DISTILLWEIGHTED.

We summarize our contributions as follows:

- We train more than 200 models across a diverse set of source and target tasks using single-source distillation, and extensively show that the choice of source model is imperative for the predictive performance of the target model. To the best of our knowledge, no previous work has addressed how to efficiently select a teacher model for (cross-domain) distillation.
- We find that *task similarity metrics* correlate well with predictive performance and can be used to efficiently select and weight source models for single- and multi-source distillation without access to any source data.
- We show that our approaches yield the best accuracy on multiple target tasks under compute and data constraints. We compare our DISTILLNEAREST and DISTILLWEIGHTED methods to two baselines (transfer learning and FixMatch), as well as the naïve case of DISTILLWEIGHTED with *equal* weighting (called DISTILLEQUAL), among others. Averaged over 8 diverse datasets, our DISTILLWEIGHTED outperforms the baselines with at least 4.5% and in particular 17.5% on CUB200.

2. Related Work

Knowledge Distillation One key aspect of our problem is to figure out how to compress single or multiple large foundation models into an efficient target model. A common approach is knowledge distillation [5, 18] where an efficient student model is trained to mimic the output of a larger teacher model. However, most single-teacher [3, 10, 11, 27, 29] or multi-teacher knowledge distillation [16, 26, 37, 44] research focuses on the closed set setup, where the teacher(s) and the student both attempts to tackle the same task. To the best of our knowledge, compressing models specializing in various tasks different from the target task has rarely been explored in the literature. Our paper explores this setup and illustrates that carefully distilling source models trained on different tasks can bring forth efficient yet accurate models.

Semi-Supervised Learning and Transfer Given our target tasks are specified in a semi-supervised setting, it is customary to review methods for semi-supervised learning (SSL). The key to SSL approaches is how to effectively propagate label information from a small labeled dataset to a large unlabeled dataset. Along this vein, methods such as pseudo-labeling/self-training [24, 42] or consistency regularization [7, 35, 38] have shown remarkable results in reducing deep networks dependencies on large labeled datasets via unlabeled data. However, most SSL approaches focus on training models from scratch without considering the availability of pre-trained models. Given the increasing availability of large pre-trained models [30, 41], recent work has started exploring the intersection between transfer learning and SSL [1, 20, 33]. However, most of these works focus on how to transfer from a single pre-trained model to the target task. Our paper, however, explores an even more practical setup: how to transfer from multiple pre-trained models to a downstream task where in-domain unlabeled data are available. In principle, we could combine our approach with a lot of previous work on SSL to (potentially) gain even larger improvements, but to keep our method simple we leave such exploration to future work and focus on how to better utilize an available set of pre-trained models.

Multi-Source Domain Adaptation Our setup also bears a resemblance with multi-source domain adaptation (MSDA) [31] in which the goal is to create a target model by leveraging multiple source models. However, MSDA methods often assume the source and target models share the same label space to perform domain alignment. We do not make such an assumption and in fact, focus on the case where the label space of source and target tasks have minimal to no overlap. Besides, a lot of the MSDA approaches [31, 43, 46, 47] rely on the availability of source data or the fact that the source and target tasks share the same model architecture to build domain invariant features. Given the discrepancy in assumptions between MSDA and our setup, we do not consider any methods from this line of work as baselines.

Transfer Learning From Multiple Sources Transfer learning from multiple different pre-trained models has been explored in different setups. Bolya et al. [9] focuses on how to select a single good pre-trained model to use as a model initialization whereas we explore how to distill an efficient

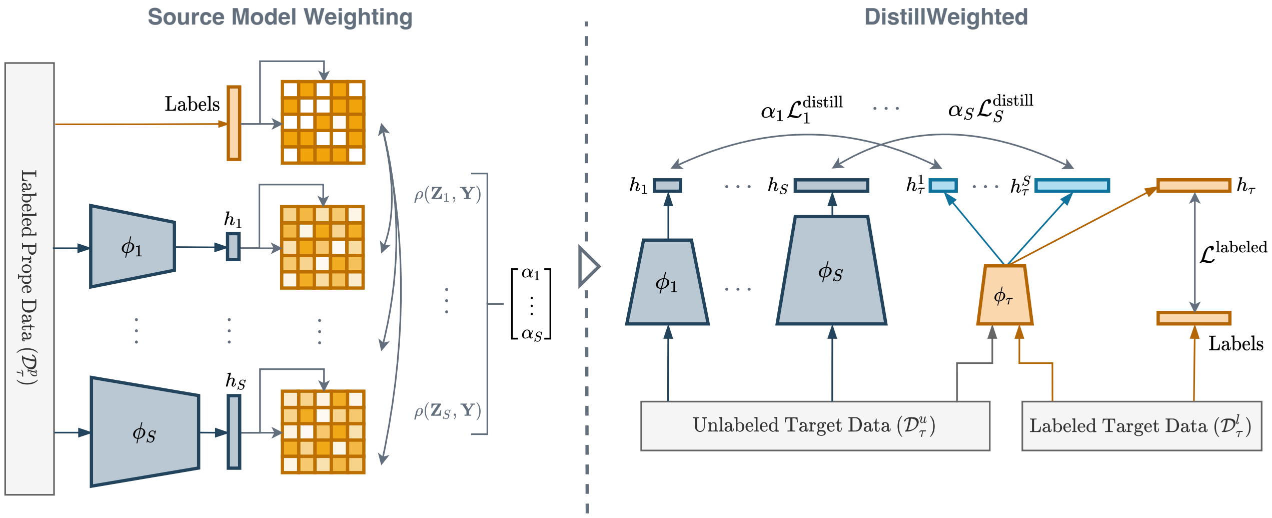
Source Model Weighting
Labels
Labeled Probe Data ($\mathcal{D}^p_\tau$)
ϕ_1
h_1
ϕ_S
h_S
$\rho(\mathbf{Z}_1, \mathbf{Y})$
$\rho(\mathbf{Z}_S, \mathbf{Y})$
α_1
α_S
DistillWeighted
$\alpha_1 \mathcal{L}_1^{\text{distill}}$ $\cdots$ $\alpha_S \mathcal{L}_S^{\text{distill}}$
h_1
h_S
h_τ^1
h_τ^S
h_τ
$\mathcal{L}^{\text{labeled}}$
ϕ_1
ϕ_S
ϕ_τ
Labels
Unlabeled Target Data ($\mathcal{D}^u_\tau$)
Labeled Target Data ($\mathcal{D}^l_\tau$)

Figure 2: We propose to weigh a set of S source models, $\mathcal{M}_s = h_s \circ \phi_s$, by using task similarity metrics to estimate the alignment of each source model with the particular target task using a small probe set of labeled data, $\mathcal{D}^p_\tau$. Since the task similarity metrics are independent of feature dimension, we can utilize source models of any architecture and from any source task. We show that by choosing the weighting, $\alpha_1, \dots, \alpha_S$, this way we are able to improve performance over transfer from ImageNet and training with FixMatch amongst others (see *e.g*. Table 1 and Figure 3).

model from the pre-trained models (i.e. our target architecture could be different from those of the source models). Agostinelli et al. [4] focuses on how to select a subset of pre-trained models to construct an (fine-tuned) ensemble, whereas we focus on creating a single model. Li et al. [25] focuses on creating a generalist representation by equally distilling multiple pre-trained models using proxy/source data (which often requires high-capacity models) whereas our goal is to construct an efficient specialist model using the target data. All these works have indicated the importance of exploring how to best leverage a large collection of pre-trained models but due to differences in setup and assumptions, we do not (and could not) compare to them.

Task Similarity / Transferability Metrics A key insight of our approach is to leverage the similarity between the target and source tasks to compare and weigh different pre-trained source models during distillation. Characterizing tasks (or similarities between tasks) is an open research question with various successes. A common approach is to embed tasks into a common vector space and characterize similarities in said space. Representative research along this line of work include Achille et al. [2], Peng et al. [32], Wallace et al. [40]. Another related line of work investigates transferability metrics [6, 9, 14, 15, 28, 39]. After all, one of the biggest use cases of task similarities is to predict how well a model transfers to new tasks. Since it is not our intention to define new task similarity/transferability metrics for distillation, we use already established metrics that capture the similarity between source representations and one-hot labels to weigh the source models. Under this purview, metrics that characterize similarities between features such as CKA [12, 22] and transferability metrics based on features [9, 14] suffice.

3. Problem Setting

The aim of this paper is to train an accurate model for a given target task, subject to limited labeled data and computational constraints (*e.g*. limited compute resources). Formally, we assume that our target task is specified via a small labeled training set $\mathcal{D}^l_\tau$. Furthermore, we assume (a) the availability of a set of unlabeled data, $\mathcal{D}^u_\tau$, associated with the target task, and (b) the ability to perform inference on a set $\mathcal{S} = \{\mathcal{M}_s\}_{s=1}^S$ of S different *source* models, $\mathcal{M}_s$, trained on various source tasks different from the target task. We emphasize that we have no access to any source data which could be practical due to storage, privacy, and computational constraints. Neither do we need full access to the source models provided we can perform inference on the models anywise (*e.g*. through an API).

We assume that the architecture of the target model, $\mathcal{M}_\tau$, must be chosen to meet any applicable computational constraints. This can imply that no suitable target architecture is available in the set of source models $\mathcal{S}$, making classical transfer learning impossible. For simplicity, we restrict our models (regardless of source or target) to classification models that can be parameterized as $\mathcal{M} = h \circ \phi$; the feature extractor ϕ embeds input $\mathbf{x}$ into a feature representation, and

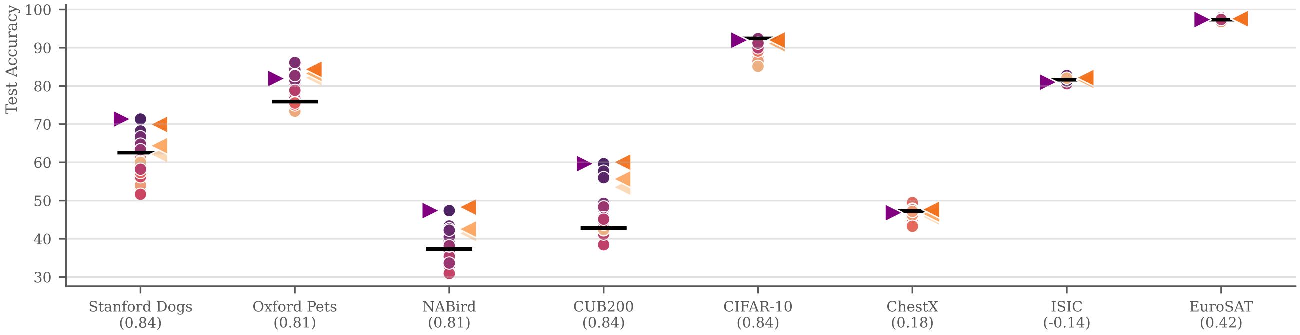

Figure 3: Test accuracy for distillation with each dot representing single-source distillation from different source models. The colors represent the task similarity for the source models (from small to large; ▨). We include the performance from fine-tuning ImageNet (——), DistillNearest; i.e. distillation of the highest ranked source model (▶) as well as DistillEqual (◀), and DistillWeighted(p) where weights are proportional to task similarity with power $p = 1$ (◀), and $p = 12$ (◀), respectively. The numbers in parentheses at the bottom are Spearman correlations between the task similarity and test accuracy for single-source distillation.

the classifier head, h, maps the feature $\phi(\mathbf{x})$ into predicted conditional class probabilities, $P(\mathbf{y} \mid \mathbf{x})$.

4. Cross-Task Distillation for Constructing Efficient Models from Foundation Models

To construct an efficient model, we propose to distill large foundation models. Along this vein, we propose two variants: (a) DistillNearest that distills the single nearest source model (Section 4.1) and (b) DistillWeighted that distills a weighted collection of source models (Section 4.2).

4.1. DistillNearest

To construct a single efficient target model, DistillNearest undergoes two steps sequentially: (a) selecting an appropriate source model and (b) distilling the knowledge from the selected source model into the target model. For ease of exposition, we start by explaining the distillation process and then discuss how to select an appropriate source model.

Distilling a selected source model. Given a selected source model $\mathcal{M}_s$, the target model $\mathcal{M}_\tau = h_\tau \circ \phi_\tau$ is trained by minimizing a weighted sum of two loss functions,

$$\mathcal{L}_{\text{single}} \overset{\text{def}}{=} \lambda \mathcal{L}^{\text{labeled}} + (1-\lambda)\mathcal{L}_s^{\text{distill}}, \tag{1}$$

where $\lambda \in [0, 1]$. The first loss function is the standard supervised objective over the labeled data,

$$\mathcal{L}^{\text{labeled}} \overset{\text{def}}{=} \frac{1}{|\mathcal{D}_\tau^l|} \sum_{(\mathbf{x}_i,\mathbf{y}_i)\in\mathcal{D}_\tau^l} \ell_{CE}\left(h_\tau(\phi_\tau(\mathbf{x}_i)), \mathbf{y}_i\right), \tag{2}$$

where $\ell_{CE}(\cdot,\cdot)$ is the cross-entropy loss. The second loss function is a distillation loss over the unlabeled data,

$$\mathcal{L}_s^{\text{distill}} \overset{\text{def}}{=} \frac{1}{|\mathcal{D}_\tau^u|} \sum_{\mathbf{x}_i\in\mathcal{D}_\tau^u} \ell_{CE}\left(h_\tau^s(\phi_\tau(\mathbf{x}_i)), \mathcal{M}_s(\mathbf{x}_i)\right). \tag{3}$$

Note, the source and target tasks do not share the same label space so we introduce an additional classifier head, h_τ^s, which maps the features from the target task feature extractor, ϕ_τ, to the label space of the source task. This additional classifier head, h_τ^s, is discarded after training and only the target classifier head, h_τ, is used for inference.

In principle, we could add additional semi-supervised losses, such as the FixMatch loss [35] to propagate label information from the labeled set to the unlabeled set for better performance, but this would add additional hyperparameters and entangle the effect of our methods. We leave such explorations to future work.

Selecting the nearest source model for distillation. Selecting a source model for distillation is an under-explored problem. Given the recent success of using task similarity metrics [9] for selecting foundation models for fine-tuning, we conjecture that high similarities between a source model and the target task could indicate better performance of the distilled model (we verify this in Section 5.2). However, quantifying similarities between tasks/models is an open research question with various successes [2, 28]. For simplicity, we pick our similarity based on one simple intuition: target examples with identical labels should have similar source representations and vice versa. Along this vein, the recently introduced metric, PARC [9] fits the bill.

For convenience, we briefly review PARC. Given a small labeled probe set $\mathcal{D}_\tau^p = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n \subseteq \mathcal{D}_\tau^l$ and a source representation of interest ϕ_s, PARC first constructs two distance matrices D_{ϕ_s}, D_Y based on the Pearson correlations between every pair of examples in the probe set;

$$\begin{aligned} D_{\phi_s} &= 1 - \mathrm{pearson}(\{\phi_s(\mathbf{x}_i)\}_{i=1}^n), \\ D_Y &= 1 - \mathrm{pearson}(\{\mathbf{y}_i\}_{i=1}^n). \end{aligned}$$

PARC is computed as the Spearman correlation between the lower triangles of the distance matrices;

$$\mathrm{PARC}(\phi_s, Y) = \mathrm{spear}\left(\{D_{\phi_s}[i,j]\}_{i<j}, \{D_Y[i,j]\}_{i<j}\right).$$

Intuitively, PARC quantifies the similarity of representations by comparing the (dis)similarity structures of examples within different feature spaces: if two representations are similar, then (dis)similar examples in one feature space should stay (dis)similar in the other feature space. In Figure 3 and 4 we show that ranking source models by PARC correlates well with test accuracy and that selecting an appropriate source model can yield significant improvements.

4.2. DistillWeighted

Above, DistillNearest assumes a single optimal source model exists for the target task, but what if no single source model aligns well with our target task? To alleviate this issue, we propose to distill multiple source models, weighted according to their similarities with the target tasks. In the following, we explain our weighted distillation objective and how the weights are constructed. Figure 2 is a schematic depiction of the approach DistillWeighted.

Weighted objective for distilling multiple sources. Given a set of source models $\mathcal{S} = \{M_s\}_{s=1}^S$, we modify the distillation loss of (1) with a weighted sum of multiple distillation losses (one for each source model):

$$\mathcal{L}_{\text{multi}} \overset{\text{def}}{=} \lambda \mathcal{L}^{\text{labeled}} + (1-\lambda)\sum_{s=1}^{S} \alpha_s \mathcal{L}_s^{\text{distill}}, \qquad (4)$$

where $\lambda, \alpha_1, \ldots, \alpha_S \in [0,1]$ ($\mathcal{L}^{\text{labeled}}$ and $\mathcal{L}_s^{\text{distill}}$ are as defined in (2) and (3), respectively). Here α_s is the relative weight assigned to each source model such that $\sum_{s=1}^S \alpha_s = 1$. Once again, we could add additional semi-supervised losses, such as the FixMatch loss, but to ensure simplicity, we leave such explorations for future research.

Task similarity weighting of source models Simply assigning equal weight to all source models is sub-optimal (e.g. weighing source models trained on ImageNet and Chest X-ray equally might not be optimal for recognizing birds). As such, we propose to compute the source weight α_s from a task similarity metric between the s-th source model and the target task. In particular, let e_s be such a similarity metric, then we compute the source weights $\{\alpha_i\}_{i\in[S]}$ as

$$\alpha_i = \frac{\underline{e}_i^p}{\sum_{s=1}^S \underline{e}_s^p}, \quad \text{where } \underline{e}_j = \max(0, e_j) \qquad (5)$$

for $j = 1, \ldots, S$. Here p is a hyperparameter to re-scale the distribution of the weights. Larger p assigns more weight to the most similar source models, while $p = 0$ corresponds to equal weights for all models (denoted DistillEqual), and $p \to \infty$ assigns all weight to the most similar source model (i.e. DistillNearest). When relevant, we use the notation DistillWeighted(p) to indicate the choice of p.

Scalability For DistillWeighted to be feasible, compared to DistillNearest, we need to ensure that the training procedure scales well with the size of $\mathcal{S}$. Since the computation of the weights $\{\alpha_s\}_{s=1}^S$ is based on the small probe set and is almost identical to the selection procedure for DistillNearest this is a negligible step. When training the target model, we merely require one forward pass on the unlabeled target dataset with each source model (to obtain pseudo-labels) as well as training of a one-layer classifier head per source model, both of which are cheap compared to the full training procedure of the target model. Nonetheless, one could employ a pre-selection of the top-k source models with the largest task similarity, thereby reducing the number of classifier heads and forward passes required. However, doing so introduces another hyperparameter, k, (i.e. how many models to use) complicating the analysis. Moreover, since large p induces such pre-selection in a *soft* manner, we leave it to future research to determine how to select the appropriate k.

5. Experiments and Results

5.1. Experimental Setup

Benchmark. Despite our methods being designed with the interest of using large vision models (that are potentially only available for inference), such a setting is intractable for our research. Thus, to allow for controlled experimentation we restrict our source models to a more tractable scale. In particular, we modify an existing transfer learning benchmark: Scalable Diverse Model Selection by [9], and use the publicly available models to construct a set of source models for each target task. Thus, we consider a set consisting of 28 models: 4 architectures (AlexNet, GoogLeNet, ResNet-18, and ResNet-50 [17, 23]) trained on 7 different source tasks (CIFAR-10, Caltech101, CUB200, NABird, Oxford Pets, Stanford Dogs, and VOC2007). For the target tasks, we consider 8 different tasks covering various image domains (Natural images: CIFAR-10, CUB200, NABird, Oxford Pets, Stanford Dogs; X-ray: ChestX; Skin Lesion Images: ISIC;

		Target Data								Oxford	Stanford	
		Labeled	Unlabeled	CIFAR-10	CUB200	ChestX	EuroSAT	ISIC	NABird	Pets	Dogs	Mean
MobileNetV3 (0.24 GFLOPs)	IN+Transfer	✓	-	92.4	42.8	47.3	97.4	81.6	37.3	75.9	62.6	67.2
	IN+FixMatch	✓	✓	**93.5**	41.9	38.5	**98.1**	**82.6**	*42.8*	83.4	*65.8*	68.3
	DISTILLRANDOMSELECTION	✓	✓	89.6	46.5	46.6	97.4	*81.8*	39.0	79.4	61.9	67.8
	(Ours) DISTILLNEAREST	✓	✓	92.0	59.6	46.8	97.4	81.0	47.4	81.9	**71.3**	72.2
	DISTILLEQUAL	✓	✓	90.8	*53.5*	45.7	97.5	81.5	41.4	*82.1*	62.1	*69.3*
	DISTILLRANDOMWEIGHTS	✓	✓	87.9	44.9	*46.9*	97.8	81.6	39.6	80.2	59.2	67.3
	(Ours) DISTILLWEIGHTED	✓	✓	*92.0*	**60.0**	**47.7**	*97.6*	82.2	**48.3**	**84.4**	69.9	**72.8**
AlexNet (0.71 GFLOPs)	IN+Transfer	✓	-	85.0	18.4	46.2	91.9	67.8	13.0	50.9	29.1	50.3
	Fine-tune Selected Source	✓	-	88.0	30.4	42.9	89.8	74.5	17.9	66.8	41.3	56.5
GoogLeNet (1.51 GFLOPs)	IN+Transfer	✓	-	91.8	42.8	41.4	96.8	80.5	36.5	84.8	65.9	67.6
	Fine-tune Selected Source	✓	-	91.6	61.2	48.6	96.9	78.3	33.0	87.8	71.8	71.2
ResNet-18 (1.83 GFLOPs)	IN+Transfer	✓	-	92.2	37.8	45.2	96.6	80.2	34.0	80.2	58.2	65.6
	Fine-tune Selected Source	✓	-	91.3	58.2	46.4	97.0	75.8	35.4	80.7	69.3	69.3
ResNet-50 (4.14 GFLOPs)	IN+Transfer	✓	-	92.9	42.0	43.4	96.8	79.9	39.9	83.3	65.9	68.0
	Fine-tune Selected Source	✓	-	93.0	70.8	43.9	97.2	81.3	47.4	84.8	79.3	74.7

Table 1: Cross-task distillation compared to baselines. MobileNetV3 models (target architecture) trained with our methods are highly competitive with baseline methods on MobileNetV3 as well as baseline methods for more demanding model architectures (source architectures: Alexnet, GoogLeNet, ResNet-18, ResNet-50). We highlight the top 3 methods, which comply with compute requirements (i.e. MobileNetV3) for each target task by **bold**, underline, and *italic*, respectively. We also indicate the target data used by different methods.

Satellite Images: EuroSAT). We treat 20% of the samples as labeled, and the remaining as unlabeled. We carefully remove any source models associated with a particular target task, if such exists, in order to avoid information leakage between source and target tasks. For the target architecture, we use MobileNetV3 [19] due to its low computational requirements compared to any of the source models. Furthermore, unless otherwise mentioned $\lambda = 0.8$ and $p = 12$. We refer the reader to the supplementary material for further details.

Baselines. We consider a set of different baselines: based on ImageNet initializations we consider IN+TRANSFER (fine-tunes ImageNet representations using only the labeled data), and IN+FIXMATCH [35] (fine-tunes the ImageNet representation using labeled and unlabeled data), and based on source model initializations we fine-tune the highest-ranked source model of each source architecture. To show the importance of using the right source model(s) to distill, we also compare DISTILLNEAREST to DISTILLRANDOMSELECTION which is the average of distilling from a randomly selected source, and for comparison to DISTILLWEIGHTED we also construct distilled models using the multi-source objective (4) with a random weight (DISTILLRANDOMWEIGHTS) and equal weights (DISTILLEQUAL). For ease of exposition, we present results for DISTILLNEAREST (Section 5.2) and DISTILLWEIGHTED (Section 5.3) in separate sections.

5.2. Results for DISTILLNEAREST

We compare DISTILLNEAREST with the baselines in Table 1 and Figure 3. Our observations are as follows.

Distillation with the right source model is better than fine-tuning from ImageNet. We observe that within the same target architecture (MobileNetv3), simply fine-tuning ImageNet representations (IN+TRANSFER) is less optimal than distilling from the most similar single model (DISTILLNEAREST). In fact, for fine-grained datasets such as CUB200, NABird, Oxford Pets, and Stanford Dogs, we observe that distilling from an appropriate source model (DISTILLNEAREST) could yield much better performance than fine-tuning from a generalist ImageNet representation. More surprisingly, even with the aid of unlabeled data, models fine-tuned from ImageNet representations using a label propagation style approach (IN+FIXMATCH) still underperform distillation-based methods by at least 3.9% on average. These observations indicate the importance of selecting the right source model for transfer/distillation.

Distilling to efficient architecture could be better than fine-tuning larger models. In Table 1, we include the performance when fine-tuning larger architectures trained on ImageNet (IN+TRANSFER) and the source model (of the same architecture) most similar to each target task selected using PARC (FINE-TUNE SELECTED SOURCE). A few observations are immediate: (a) our choice of task similarity metric is effective for transfer; across all 4 architectures, we observe at least 4% improvement over simple fine-tuning from ImageNet, which validates the results by Bolya et al. [9], and (b) with the aid of unlabeled data and distillation, the computationally efficient architecture MobileNetV3 can outperform larger architectures fine-tuned on labeled data from the target task (i.e. AlexNet, GoogLeNet, ResNet-18). Although underperforming fine-tuning a ResNet-50 initialized

		CIFAR-10	CUB200	ChestX	EuroSAT	ISIC	NABird	Oxford Pets	Stanford Dogs	Mean
Pseudo	CKA	0.72	0.62	0.23	0.39	-0.04	0.31	0.69	0.11	0.38
	PARC	0.79	0.79	0.02	0.17	0.06	0.48	0.72	0.54	0.45
	RSA	0.82	0.31	-0.11	0.30	**0.10**	-0.03	0.65	0.38	0.30
Feature	CKA	0.82	0.39	**0.36**	0.21	-0.04	0.47	0.69	0.55	0.43
	PARC	0.84	**0.84**	0.18	**0.42**	-0.14	**0.81**	0.81	0.84	**0.58**
	RSA	**0.86**	0.81	0.03	0.38	0.03	0.28	**0.89**	**0.85**	0.52

Table 2: Spearman correlation between test accuracy after all possible single-source distillations and task similarities associated with the source models. Generally feature representations correlate better with distillation performance compared to pseudo-label representations.

		CIFAR-10	CUB200	ChestX	EuroSAT	ISIC	NABird	Oxford Pets	Stanford Dogs	Mean
Pseudo	CKA	99.1	95.6	97.4	99.6	98.8	89.4	**100.0**	97.6	97.2
	PARC	99.5	**100.0**	95.5	99.6	98.5	99.7	98.8	**99.7**	98.9
	RSA	**100.0**	77.7	96.5	99.7	98.5	87.2	98.6	97.6	94.5
Feature	CKA	**100.0**	95.6	97.0	**99.8**	**99.0**	93.3	**100.0**	96.4	97.6
	PARC	**100.0**	**100.0**	**97.8**	99.7	98.3	**100.0**	97.1	98.5	**98.9**
	RSA	**100.0**	**100.0**	96.7	99.8	98.9	94.9	98.9	98.8	98.5

Table 3: Relative accuracy of top-3 single-source distilled models selected by task similarity over the average of the 3 actual best models. We compute the average test accuracy of the top-3 highest ranked target models and divide it by the average of the 3 actually best-performing target models.

with the most similar ResNet-50 source model by a mere average of 2.5%-points (FINE-TUNE SELECTED SOURCE), using a ResNet-50 would require 17.5× more computations during inference to achieve such improvements.

5.2.1 Task Similarity Metrics for DISTILLNEAREST

One key component of DISTILLNEAREST is to select the source model to perform cross-task distillation on using task similarity metrics. Despite many many existing metrics for quantifying task similarities, their effectiveness for distillation remains unclear. Given the myriads of metrics, we restrict our focus to metrics that can capture similarities between a source representation of a target example and its one-hot label representation. Along this vein, two questions arise: which metric to use for comparing representations, and which representations from a source model should be used to represent a target example?

For the first question, we look into multiple metrics in the literature that compares various representations: CKA [12], RSA [14], and PARC [9]. For the second question, we look into the common representations from a source model: the features ϕ and the probabilistic outputs $h \circ \phi$.

To establish the effectiveness of our choice of similarity metric, we report the Spearman correlation between the task similarities and the test accuracy of the distilled models in Table 2. We see that features from the source models can better capture the correlation between the source models and the test accuracy of the distilled models, than the probabilistic pseudo-labels. In addition, we also see a much higher correlation among natural tasks (compared to specialized tasks such as ChestX, EuroSAT, and ISIC) which suggests that our choice of task similarity is effective at selecting similar tasks. Besides, we also observe a higher correlation when using PARC compared to the other metrics, thus validating our choice of using PARC as the default metric.

To further establish the effectiveness of our metrics to rank various source models, we compute the relative test accuracy between the top-3 models most similar to the target task and the top-3 best models after distillation (see Table 3). Again, we observe that all three metrics are capable of ranking affinity between source models, but ranking the models with PARC outperforms the other two metrics.

5.3. Results for DISTILLWEIGHTED

From Table 1, we observe that DISTILLWEIGHTED compares favorably to DISTILLNEAREST, thus the conclusions for DISTILLNEAREST translates to DISTILLWEIGHTED. Yet, one particular task, Oxford Pets, is worth more attention. On Oxford Pets (classification of different breeds of cats and dogs), we observe that distilling from multiple weighted sources (DISTILLWEIGHTED) is much better than distilling from the single most similar source (DISTILLNEAREST), which is a ResNet-18 trained on Caltech101 (that can recognize concepts such as Dalmatian dog, spotted cats, etc.). Although the most similar source model contains relevant information for recognizing different breeds of dogs and cats, it might not contain all relevant knowledge from the set of source models that could be conducive to recognizing all visual concepts in Oxford Pets. In fact, we observe that the second most similar model is a GoogLeNet model trained on Stanford Dogs to recognize more dog breeds than the most similar source model (but incapable of recognizing cats). In this case, DISTILLWEIGHTED allows aggregation of knowledge from multiple sources and can effectively combine knowledge from different source models for a more accurate target model than distillation from a single source. This suggests that *under certain conditions such as high heterogeneity in data, distilling from multiple source models can outperform distilling a single best source model.*

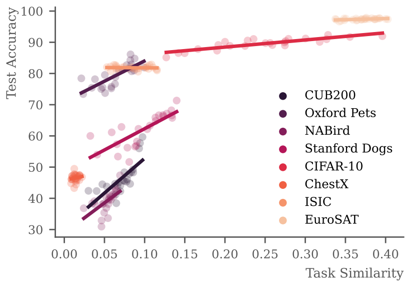

Figure 4: Test accuracy of single-source distillation and raw task similarity score using PARC on the feature representations. The scores are on different scales for different tasks, but almost all tasks have a positive correlation between test accuracy and task similarity.

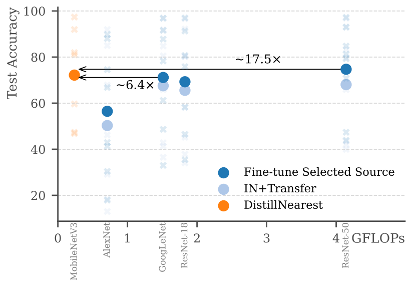

Figure 5: Average test accuracy over the 8 target tasks vs. compute requirements for a single forward pass at inference. Using DISTILLNEAREST with an efficient target architecture outperforms (or is comparable to) fine-tuning larger models.

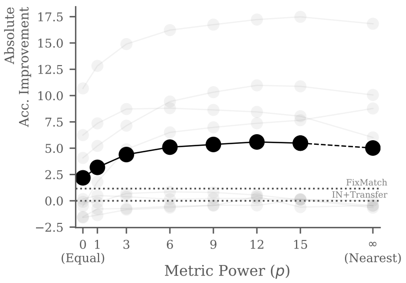

Figure 6: Improvement over IN+TRANSFER. Here ● is the average improvement over all eight target tasks and ● represents the performance on a target task. Note, $p = 0$ corresponds to DISTILLEQUAL, and $p = \infty$ corresponds to DISTILLNEAREST.

5.3.1 Task Similarity Metrics for Weighing Sources

We have established that our task similarity metric can capture the correlation between the source model representations and the test accuracy of the distilled models. However, it is not a priori clear that weighing source models based on the ranking of their affinity to the target task would yield better performance for multi-source distillation. As such, we investigate alternative choices of weighing schemes for a subset of 5 target tasks (CUB200, EuroSAT, ISIC, Oxford Pets, Stanford Dogs): INVERSE (weights are inversely proportional to task similarity), DISTILLRANDOMWEIGHTS (weights are sampled uniformly on a 4-simplex), DISTILLRANDOMSELECTION (randomly selecting a single source model), and DISTILLEQUAL (equal weights for all models).

Through Figure 1, we find that distilling from a single or set of source models ranked using the similarity metric is much more effective than distilling from source models that are weighted randomly or equally (DISTILLRANDOMWEIGHTS or DISTILLEQUAL). In addition, the fact that INVERSE underperforms IN+TRANSFER on average suggests that it is crucial to follow the ranking induced by the similarity metrics when distilling the sources and that the metric ranks both the most similar source models and the least similar source models appropriately.

5.3.2 Effect of p

Our task similarity metrics give a good ranking of which source models to select for distillation but it is unclear whether the similarity score could be used directly without any post-processing. To investigate, we visualize the relationship between the test accuracy of the models distilled from a single source and our task similarity. From Figure 4, it is clear that the distribution of task similarities depends on the target task, which motivates our normalization scheme.

In addition, it is not apriori clear that the weights should scale linearly with the similarity scores. Thus, we investigate the effect of the rescaling factor, p, for constructing the weights. In Figure 6, we see that although no rescaling ($p = 1$) outperforms equal weighting, it is less optimal than *e.g.* $p = 12$ (our default). This suggests that task similarity and good weights have a monotonic, but non-linear relationship.

5.4. Additional Ablations and Analyses

Due to space constraints, we include additional ablations and analyses in the supplementary materials. We summarize the main findings as follows.

ResNet-50 as target model. Averaged over 8 tasks, DISTILLWEIGHTED outperforms both IN+TRANSFER and DISTILLEQUAL by 5.6% and 3.8%, respectively. Also, compared to ImageNet initialization, using DISTILLWEIGHTED with the most similar ResNet-50 source model as target model initialization improves accuracy by 1.0%.

Improvements on VTAB. DISTILLWEIGHTED outperforms IN+TRANSFER averaged over the ● *Natural* and ● *Specialized* tasks of VTAB, by 5.1% and 0.8%, respectively. DISTILLNEAREST outperform by 4.8% and 0.6%, respectively.

Fewer labels. DISTILLWEIGHTED and DISTILLNEAREST outperform IN+TRANSFER (by 6.8% and 4.4%, respectively) under a setup with even fewer labeled samples.

Additional analysis of task similarity metrics. We consider additional correlation metrics and top-k relative accuracies of the selected models — all supporting the usefulness of task similarity to weigh and select source models.

6. Conclusion

We investigate the use of diverse source models to obtain efficient and accurate models for visual recognition with limited labeled data. In particular, we propose to distill multiple diverse source models from different domains weighted by their relevance to the target task without access to any source data. We show that under computational constraints and averaged over a diverse set of target tasks, our methods outperform both transfer learning from ImageNet initializations and state-of-the-art semi-supervised techniques.

Acknowledgments This work was funded in part by NSF 2144117, and in part by Aarhus University Centre for Digitalisation, Big Data and Data Analytics (DIGIT).

References

[1] A. Abuduweili, X. Li, H. Shi, C.-Z. Xu, and D. Dou. Adaptive consistency regularization for semi-supervised transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6923–6932, 2021. 2

[2] A. Achille, M. Lam, R. Tewari, A. Ravichandran, S. Maji, C. Fowlkes, S. Soatto, and P. Perona. Task2Vec: Task embedding for meta-learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6429–6438, 2019. ISBN 9781728148038. doi: 10.1109/ICCV.2019.00653. 3, 4

[3] R. Adriana, B. Nicolas, K. S. Ebrahimi, C. Antoine, G. Carlo, and B. Yoshua. Fitnets: Hints for thin deep nets. *Proc. ICLR*, 2, 2015. 2

[4] A. Agostinelli, J. Uijlings, T. Mensink, and V. Ferrari. Transferability Metrics for Selecting Source Model Ensembles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7936–7946, 2022. URL http://arxiv.org/abs/2111.13011. 3

[5] L. J. Ba and R. Caruana. Do Deep Nets Really Need to be Deep? *Advances in Neural Information Processing Systems*, 3(January):2654–2662, 2014. ISSN 10495258. 2

[6] Y. Bao, Y. Li, S.-L. Huang, L. Zhang, L. Zheng, A. Zamir, and L. Guibas. An information-theoretic approach to transferability in task transfer learning. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2309–2313. IEEE, 2019. 3

[7] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019. 2

[8] E. Bisong and E. Bisong. Google colaboratory. *Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners*, pages 59–64, 2019. 1

[9] D. Bolya, R. Mittapalli, and J. Hoffman. Scalable Diverse Model Selection for Accessible Transfer Learning, 2021. ISSN 10495258. 2, 3, 4, 5, 6, 7

[10] K. Borup and L. N. Andersen. Even your teacher needs guidance: Ground-truth targets dampen regularization imposed by self-distillation. *Advances in Neural Information Processing Systems*, 34:5316–5327, 2021. 2

[11] J. H. Cho and B. Hariharan. On the Efficacy of Knowledge Distillation. *ICCV*, 2019. ISSN 00262714. doi: 10.1016/0026-2714(74)90354-0. 2

[12] C. Cortes, M. Mohri, and A. Rostamizadeh. Algorithms for learning kernels based on centered alignment. *Journal of Machine Learning Research*, 13:795–828, 2012. ISSN 15324435. 3, 7

[13] M. Dehghani, J. Djolonga, B. Mustafa, P. Padlewski, J. Heek, J. Gilmer, A. Steiner, M. Caron, R. Geirhos, I. Alabdulmohsin, R. Jenatton, L. Beyer, M. Tschannen, A. Arnab, X. Wang, C. Riquelme, M. Minderer, J. Puigcerver, U. Evci, M. Kumar, S. van Steenkiste, G. F. Elsayed, A. Mahendran, F. Yu, A. Oliver, F. Huot, J. Bastings, M. P. Collier, A. Gritsenko, V. Birodkar, C. Vasconcelos, Y. Tay, T. Mensink, A. Kolesnikov, F. Pavetić, D. Tran, T. Kipf, M. Lučić, X. Zhai, D. Keysers, J. Harmsen, and N. Houlsby. Scaling Vision Transformers to 22 Billion Parameters. 2023. 1

[14] K. Dwivedi and G. Roig. Representation similarity analysis for efficient task taxonomy & transfer learning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:12379–12388, 2019. ISSN 10636919. doi: 10.1109/CVPR.2019.01267. 3, 7

[15] K. Dwivedi, J. Huang, R. M. Cichy, and G. Roig. Duality diagram similarity: a generic framework for initialization selection in task transfer learning. In *European Conference on Computer Vision*, pages 497–513. Springer, 2020. 3

[16] T. Fukuda, M. Suzuki, G. Kurata, S. Thomas, J. Cui, and B. Ramabhadran. Efficient knowledge distillation from an ensemble of teachers. In *Interspeech*, pages 3697–3701, 2017. 2

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem:770–778, 2016. ISSN 10636919. doi: 10.1109/CVPR.2016.90. 5

[18] G. Hinton, O. Vinyals, and J. Dean. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*, pages 1–9, 2015. URL http://arxiv.org/abs/1503.02531. 2

[19] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, and L. H. Adam1. Searching for MobileNetV3. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-Octob:1314–1324, 2019. ISSN 15505499. doi: 10.1109/ICCV.2019.00140. 6

[20] A. Islam, C.-F. R. Chen, R. Panda, L. Karlinsky, R. Feris, and R. J. Radke. Dynamic Distillation Network for Cross-Domain Few-Shot Recognition with Unlabeled Data. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 3584–3595. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/1d6408264d31d453d556c60fe7d0459e-Paper.pdfhttp://arxiv.org/abs/2106.07807. 2

[21] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby. Big Transfer (BiT): General Visual Representation Learning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12350 LNCS:491–507, 2020. ISSN 16113349. doi: 10.1007/978-3-030-58558-7{_}29. 1

[22] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019. 3

[23] A. Krizhevsky, I. Sutskever, G. E. Hinton, S. Ilya, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. 5

[24] D.-H. Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, page 896, 2013. 2

[25] Z. Li, A. Ravichandran, C. Fowlkes, M. Polito, R. Bhotika, and S. Soatto. Representation Consolidation for Training Expert Students, 2021. URL http://arxiv.org/abs/2107.08039. 3

[26] Y. Liu, W. Zhang, and J. Wang. Adaptive multi-teacher multi-level knowledge distillation. *Neurocomputing*, 415:106–113, 2020. 2

[27] S.-I. Mirzadeh, M. Farajtabar, A. Li, H. Ghasemzadeh, N. Levine, A. Matsukawa, and H. Ghasemzadeh. Improved Knowledge Distillation via Teacher Assistant: Bridging the Gap Between Student and Teacher. *arXiv preprint arXiv:1902.03393*, 2019. URL http://arxiv.org/abs/1902.03393. 2

[28] C. Nguyen, T. Hassner, M. Seeger, and C. Archambeau. Leep: A new measure to evaluate transferability of learned representations. In *International Conference on Machine Learning*, pages 7294–7305. PMLR, 2020. 3, 4

[29] W. Park, D. Kim, Y. Lu, and M. Cho. Relational Knowledge Distillation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019. ISBN 9781728132938. doi: 10.1109/CVPR.2019.00409. 2

[30] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2

[31] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 2

[32] X. Peng, Y. Li, and K. Saenko. Domain2Vec: Domain Embedding for Unsupervised Domain Adaptation. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, editors, *Computer Vision – ECCV 2020*, pages 756–774, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58539-6. 3

[33] C. P. Phoo and B. Hariharan. Self-training For Few-shot Transfer Across Extreme Task Differences. In *International Conference on Learning Representations*, pages 1–19, 2021. URL http://arxiv.org/abs/2010.07734https://openreview.net/forum?id=O3Y56aqpChA. 2

[34] A. Rekognition. URL https://aws.amazon.com/rekognition/. 1

[35] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, C.-L. L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. A. Raffel. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 596–608. Curran Associates, Inc., 2020. 2, 4, 6

[36] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. 1

[37] X. Tan, Y. Ren, D. He, T. Qin, Z. Zhao, and T.-Y. Liu. Multilingual neural machine translation with knowledge distillation. *arXiv preprint arXiv:1902.10461*, 2019. 2

[38] A. Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 2

[39] A. T. Tran, C. V. Nguyen, and T. Hassner. Transferability and hardness of supervised classification tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1395–1405, 2019. 3

[40] B. Wallace, Z. Wu, and B. Hariharan. Can we characterize tasks without labels or features? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1245–1254, June 2021. 3

[41] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019. 2

[42] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020. 2

[43] R. Xu, Z. Chen, W. Zuo, J. Yan, and L. Lin. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3964–3973, 2018. 2

[44] S. You, C. Xu, C. Xu, and D. Tao. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1285–1294, 2017. 2

[45] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer. Scaling Vision Transformers. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022-June:12094–12103, 2022. ISSN 10636919. doi:

10.1109/CVPR52688.2022.01179. 1

[46] H. Zhao, S. Zhang, G. Wu, J. M. Moura, J. P. Costeira, and G. J. Gordon. Adversarial multiple source domain adaptation. *Advances in neural information processing systems*, 31, 2018. 2

[47] S. Zhao, G. Wang, S. Zhang, Y. Gu, Y. Li, Z. Song, P. Xu, R. Hu, H. Chai, and K. Keutzer. Multi-source distilling domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12975–12983, 2020. 2