Supplementary Materials for 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

Code: github.com/Kennethborup/DistillWeighted

We present additional results and implementation details in the supplementary. To avoid confusion, we use the same set of index numbers as in the main text to refer to the tables and figures. Please find Tables 1-3 and Figures 1-6 in the main text.

A. Additional Ablations and Analyses

A.1. Results on VTAB

We report the results of our VTAB [7] experiment in Table 4. On VTAB, We find that both DISTILLWEIGHTED and DISTILLNEAREST distillation outperform IN+TRANSFER on each of the ● *Natural* tasks. Particularly, DISTILLWEIGHTED outperforms IN+TRANSFER with 13.9%-points on CIFAR-10 and 10.6%-points on Sun397 and averaged across ● *Natural* DISTILLWEIGHTED outperforms IN+TRANSFER with 5.1%-points. Average over ● *Specialized* both DISTILLWEIGHTED and DISTILLNEAREST outperform IN+TRANSFER, although with a small margin. Finally, averaged over ● *Structured* IN+TRANSFER outperforms our methods, but due to the nature of these tasks, we do not expect source models to transfer well to these tasks.[1] Yet, we still obtain the best accuracy on DMLab, dSpr-Loc, and sNORB-Azimuth.

A.2. Relative accuracy of single-source distillation

Similarly to Table 3, we extend our evaluation of how well the task similarity selects the best source models for single-source distillation. We report the ratio between the average test accuracy of the top-k target models ranked using the task similarity and the average test accuracy for the actual top-k target models found after the fact in Table 5, Table 6, and Table 7 for $k = 1$, $k = 3$, and $k = 5$, respectively.

We find that generally, using task similarity on feature representations rather than the corresponding pseudo-labels yields better rankings, but also that PARC shows very little difference between features and pseudo-labels for all considered $k \in \{1, 3, 5\}$.

Relative accuracy over all k. The relative accuracy measure reported above is sensitive to k and the actual accuracy values of the models. I.e. if a metric flips the order of the best and second best model when there is a notable performance gap between the two models, the relative accuracy for $k = 1$ will be low, and we might be mistaken to believe the metric is not working well. However, the metric might rank every model for $k > 2$ perfectly correct, and since we typically utilize the full set of source models, the initial mistake should not be detrimental to the selection of the task similarity metric. Thus, in Figure 7 we plot the relative accuracy for each task similarity metric and all $k \in \{1, \dots, S\}$. We find that while PARC on feature representations is outperformed by both PARC and CKA on pseudo-labels for $k < 3$, PARC on feature representations outperforms all the other metrics for $k \geq 3$. In particular, from Table 8 we have that on average over all $k < S$, PARC, performs the best.

A.3. Ablation of p for DISTILLWEIGHTED

We report the values associated with Figure 6 for each target task and all considered choices of p in Table 9.

A.4. DISTILLWEIGHTED with ResNet-50 as target architecture

In the main part of the article, we consider the computationally constrained setting, where some compute budget restricts the possible size of our target model. Thus, we use MobileNetV3 models as target models throughout the main paper. However, in Table 10 we remove the computational budget and allow the target model to be of any architecture, and particularly we use a ResNet-50 as the target model.

We compare DISTILLWEIGHTED (with $p = 0$ and $p = 12$) initialized with either ImageNet pre-trained weights

[1] The ● *Structured* tasks are mainly (ordinal) regression tasks transformed into classification tasks, and thus it seems reasonable to expect very general features (such as those from an ImageNet pre-trained model) to generalize better to such constructed tasks than specialized source models.

	Caltech101	CIFAR-100	DTD	Flowers102	Pets	SVHN	Sun397	Natural	Camelyon	EuroSAT	Resisc45	Retinopathy	Specialized	Clevr-Count	Clevr-Dist	DMLab	KITTI-Dist	dSpr-Loc	dSpr-Ori	sNORB-Azim	sNORB-Elev	Structured	Mean
IN+Transfer	88.1	47.0	57.4	85.8	82.8	75.3	27.8	66.3	**81.0**	95.0	80.0	**72.7**	82.2	**73.1**	**55.9**	43.6	**75.7**	18.7	**58.6**	21.2	**46.0**	**49.1**	**62.4**
DISTILLWEIGHTED	88.6	**60.9**	**62.4**	86.1	84.4	79.0	**38.4**	**71.4**	80.6	**95.9**	**83.3**	72.2	**83.0**	57.4	45.6	44.6	67.7	**27.4**	44.9	23.9	38.2	43.7	62.2
DISTILLNEAREST	**88.9**	59.5	61.9	**86.2**	**84.5**	**79.5**	37.6	71.1	80.5	95.8	83.2	71.7	82.8	60.5	45.4	**45.2**	67.9	20.8	40.6	**24.2**	36.5	42.6	61.6

Table 4: Top-1 accuracy by dataset in VTAB. The accuracy for each task is in grey, and the average accuracy for each category of tasks is in black. Note, the • *Mean* is the average across all tasks, not categories. The largest value in each column is marked in bold. Here DISTILLWEIGHTED is with $p = 9$.

		CIFAR-10	CUB200	ChestX	EuroSAT	ISIC	NABird	Oxford Pets	Stanford Dogs	Mean
Pseudo	CKA	**99.6**	**100.0**	**96.1**	99.5	98.1	**100.0**	**100.0**	**100.0**	**99.2**
	PARC	99.3	**100.0**	93.6	99.5	**98.3**	**100.0**	98.4	**100.0**	98.6
	RSA	99.3	74.8	94.8	99.5	**98.3**	86.6	97.8	95.6	93.4
Feature	CKA	**99.6**	81.0	92.6	**99.8**	**98.3**	**100.0**	**100.0**	**100.0**	96.4
	PARC	**99.6**	**100.0**	94.6	99.5	97.7	**100.0**	95.1	**100.0**	98.3
	RSA	**99.6**	**100.0**	92.6	99.5	**98.3**	80.6	**100.0**	**100.0**	96.3

Table 5: Relative accuracy of top-1 single-source distilled target model selected by task similarity over the best model found in hindsight. We compute the test accuracy of the highest-ranked target model (ranked by some task similarity) and divide this by the test accuracy of the best-performing target model.

		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 6: (Identical to Table 3) Relative accuracy of top-3 single-source distilled target models selected by task similarity over the average of the 3 best models found in hindsight. We compute the average test accuracy of the top-3 highest ranked target models and divide this average by the average test accuracy of the 3 best-performing target models.

or the weights of the highest ranked ResNet-50 source model to IN+TRANSFER and FINE-TUNE SELECTED SOURCE. We find that DISTILLWEIGHTED initialized from ImageNet outperforms IN+TRANSFER on average for both equal weighting and $p = 12$, but underperforms FINE-TUNE SELECTED SOURCE for both p. However, since FINE-TUNE SELECTED SOURCE is initialized from well-selected source model weights, the comparison is not entirely fair. Thus, we also consider the case where we initialize the target model for DISTILLWEIGHTED with the weights of the highest ranked ResNet-50 source model, and find that for $p = 12$ DISTILLWEIGHTED performs on par with FINE-TUNE SELECTED SOURCE.

		CIFAR-10	CUB200	ChestX	EuroSAT	ISIC	NABird	Oxford Pets	Stanford Dogs	Mean
Pseudo	CKA	99.3	98.7	**98.3**	99.7	99.0	92.9	99.2	98.4	98.2
	PARC	99.7	**100.0**	96.7	99.7	98.9	94.5	**99.4**	98.4	98.4
	RSA	99.7	83.2	97.6	99.8	99.0	84.9	99.2	92.8	94.5
Feature	CKA	**99.7**	97.4	97.7	99.8	98.9	96.5	99.2	97.8	98.4
	PARC	**99.7**	100.0	97.9	99.8	99.1	**99.7**	97.5	**99.7**	**99.2**
	RSA	**99.7**	99.7	97.9	**99.8**	**99.2**	97.9	98.9	**99.7**	99.1

Table 7: Relative accuracy of top-5 single-source distilled target models selected by task similarity over the average of the 5 best models found in hindsight. We compute the results analogously to Table 6 with $k = 5$.

	CKA	PARC	RSA
Pseudo	0.985	0.990	0.974
Feature	0.986	**0.993**	0.991

Table 8: The mean relative accuracy, across all k, for each metric in Figure 7. The average is bounded in $(0, 1]$, and 1 corresponds to perfect ordering by task similarity. We find that using feature representations consistently outperforms pseudo-labels and that for both feature representations and pseudo-labels PARC performs the best.

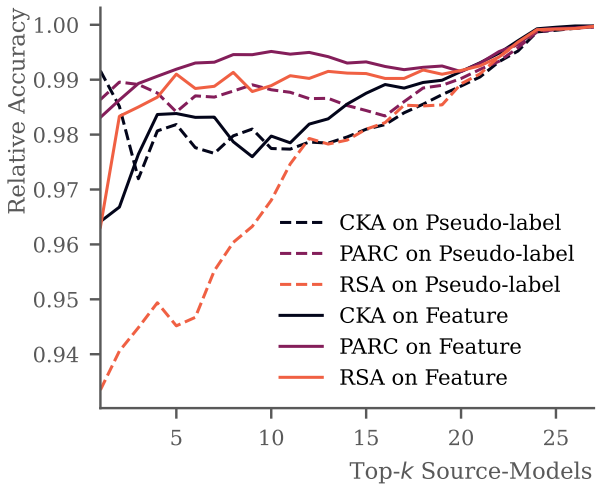

Figure 7: Relative accuracy of top-k single-source distilled target models selected by task similarity over the average of the top-k actual best target models found in hindsight. If the ordering by task similarity were perfectly correct, the relative accuracy would be 1 for all k. See Table 8 for the average of each metric across all k.

	CIFAR-10	CUB200	ChestX	EuroSAT	ISIC	NABird	Oxford Pets	Stanford Dogs	Mean
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
DISTILLEQUAL	90.8	53.5	45.7	97.5	81.5	41.4	82.1	62.1	69.3
DISTILLWEIGHTED(1)	91.1	55.6	46.5	97.9	81.5	42.5	83.3	64.4	70.3
DISTILLWEIGHTED(3)	91.6	57.7	46.5	97.7	82.3	44.5	84.6	67.4	71.6
DISTILLWEIGHTED(6)	91.8	59.0	46.7	97.5	82.5	46.7	**84.7**	69.1	72.3
DISTILLWEIGHTED(9)	92.0	59.6	46.8	97.6	82.4	47.6	84.5	69.5	72.5
DISTILLWEIGHTED(12)	92.0	60.0	**47.7**	97.6	82.2	**48.3**	84.4	69.9	**72.8**
DISTILLWEIGHTED(15)	92.6	**60.3**	46.7	97.5	81.7	48.2	83.9	70.2	72.6
DISTILLNEAREST	92.0	59.6	46.8	97.4	81.0	47.4	81.9	**71.3**	72.2

Table 9: Test accuracy of DISTILLWEIGHTED with various choices of p, compared to the baseline methods of IN+TRANSFER and IN+FIXMATCH. We highlight the largest value for each target task in **bold**, and the results are also visualized in Figure 6.

A.5. Normalization of task similarity for source model weighting

We propose to choose the weights $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_S)$ as

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

for $j = 1, \dots, S$, and e_s is the task similarity for source model $\mathcal{M}_s$, evaluated on the target task, normalized to satisfy $e_s \in [0, 1]$ with min-max normalization over all e_s.

	Model Init.	CIFAR-10	CUB200	ChestX	EuroSAT	ISIC	NABird	Oxford Pets	Stanford Dogs	Mean
IN+Transfer	ImageNet	92.9	42.0	43.4	96.8	79.9	39.9	83.3	65.9	68.0
Fine-tune Source	Source	**93.0**	**70.8**	43.9	97.2	**81.3**	47.4	84.8	**79.3**	**74.7**
DISTILLEQUAL	ImageNet	87.8	57.3	46.1	97.0	78.9	42.4	84.1	64.5	69.8
DISTILLWEIGHTED(12)	ImageNet	91.5	64.5	45.4	97.0	78.9	49.8	**87.1**	74.2	73.6
DISTILLEQUAL	Source	87.5	68.8	45.5	**97.4**	81.2	43.2	81.9	65.1	71.3
DISTILLWEIGHTED(12)	Source	91.6	70.0	**47.6**	97.0	80.8	**50.0**	85.7	73.8	74.6

Table 10: DISTILLWEIGHTED with ResNet-50 as target model architecture. We compare fine-tuning of the highest ranked source model [2] with DISTILLWEIGHTED to both ImageNet-initialized target models and target models initialized from the highest ranked ResNet-50 source model. For $p = 12$, DISTILLWEIGHTED performs on par with fine-tuning the selected source model. The largest value for each target task is in **bold**.

Here, the hyperparameter, p can be used to increase/decrease the relative weight on the highest ranked source models, with the extremes $p = 0$ and $p \to \infty$ corresponding to equal weight and single-source distillation, respectively. An alternative way to obtain our normalization is to use the softmax function on the task similarities,

$$\alpha_i = \frac{\exp\left(\frac{e_i}{T}\right)}{\sum_{s=1}^{S} \exp\left(\frac{e_s}{T}\right)}.$$

This does not require clipping the task similarity at 0, and with the temperature, T, we can adjust the relative weight on particular source models. Here, large T flattens the weights, and $T \to \infty$ corresponds to an equal weighting of all source models, while small T increases the weight on the highest-ranked source models. Quantitatively, the two normalization methods can yield similar transformations with appropriate choices of p and T - see Figure 8.

A.6. Smaller amount of labeled data

We now repeat the experiment of the main paper across the 8 target datasets with a reduced amount of labeled samples. Here, we reduce the number of labeled samples to 5% (rather than 20%) of the training set and report the accuracy in Table 11. We find a similar pattern as observed in the main experiment, where DISTILLWEIGHTED distillation on average outperforms IN+TRANSFER irrespective of the choice of p. For $p = 9$ DISTILLWEIGHTED outperforms IN+TRANSFER by 6.8%-point on average and in particular 15.5%-points on CUB200, whereas the only loss in performance is on ChestX with a drop of 0.9%-point.

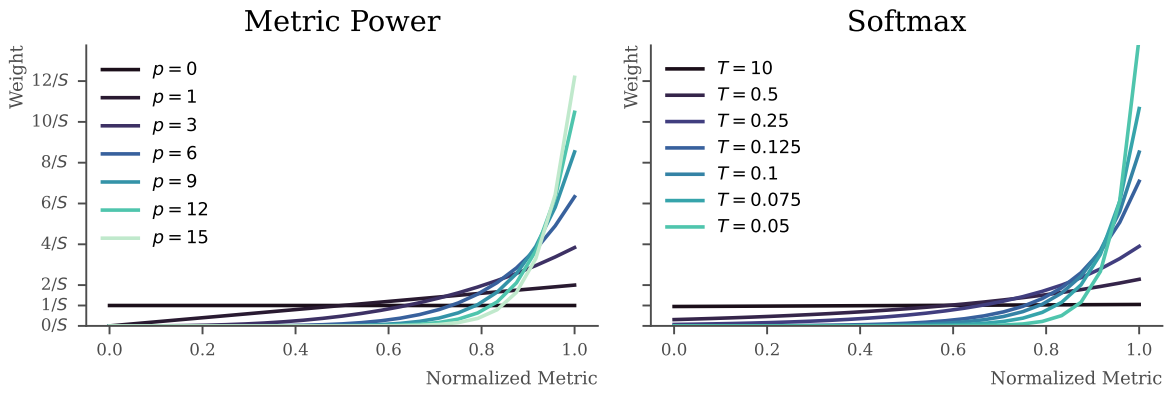

Figure 8: Transformation of weights for various choices of power (left) or softmax temperature (right). Here S is the number of source models, and we consider equidistantly distributed normalized metrics.

	CIFAR-10	CUB200	ChestX	EuroSAT	ISIC	NABird	Oxford Pets	Stanford Dogs	Mean
IN+Transfer	88.0	16.8	**43.5**	94.8	73.9	14.4	55.0	38.9	53.2
DISTILLWEIGHTED(1)	88.1	29.2	42.3	**95.9**	76.3	20.5	66.6	42.1	57.6
DISTILLWEIGHTED(9)	**90.2**	**32.3**	42.6	95.9	**76.7**	**24.8**	**68.2**	49.0	**60.0**
DISTILLNEAREST	87.2	31.4	39.7	95.1	75.4	24.0	58.9	**49.7**	57.7

Table 11: Distillation on the eight target tasks with only 5% labeled samples per task. Again, we compare to the baseline of IN+TRANSFER. The largest value for each target task is in **bold**.

A.7. Different Measures of Correlation

In order to evaluate the quality of a task similarity metric to estimate the performance of a target model after distillation, we consider the correlation between the computed metric and the actual observed performance after distillation. However, since we have no reason to believe that the relationship is linear, we consider the Spearman correlation in the main paper. However, for completeness of exposition, we report Pearson correlation and Kendall's Tau in Table 12 and Table 13, respectively. For both these correlation measures, the overall conclusions are the same: Using feature representations is preferable to pseudo-labels, and PARC generally outperforms both CKA and RSA, albeit not by much over CKA.

A.8. Choice of Task Similarity Metrics

Recently, multiple measures intended to estimate the transferability of a source model have been proposed. However, despite the very recently published Multi-Source Leep (MS-LEEP) and Ensemble Leep (E-Leep) no task similarity metric considers the estimation over multiple models at once

		CIFAR-10	CUB200	ChestX	EuroSAT	ISIC	NABird	Oxford Pets	Stanford Dogs	Mean
Pseudo	CKA	0.62	**0.85**	0.07	0.30	-0.06	0.33	0.67	0.21	0.37
	PARC	0.75	0.74	-0.03	0.27	-0.00	0.36	0.63	0.51	0.40
	RSA	0.75	0.13	-0.07	0.38	0.04	-0.09	0.66	0.40	0.27
Feature	CKA	0.84	0.60	**0.39**	0.29	0.00	0.30	0.71	0.54	0.46
	PARC	0.86	0.73	0.17	**0.46**	-0.06	**0.58**	0.77	0.78	**0.54**
	RSA	**0.90**	0.85	0.07	0.45	**0.04**	0.27	**0.87**	**0.83**	0.54

Table 12: Pearson correlation between test accuracy after all possible single-source distillations and task similarity associated with the source models. Similar to Table 2.

		CIFAR-10	CUB200	ChestX	EuroSAT	ISIC	NABird	Oxford Pets	Stanford Dogs	Mean
Pseudo	CKA	0.51	0.46	0.16	0.28	-0.05	0.24	0.49	0.07	0.27
	PARC	0.61	0.64	0.01	0.12	0.02	0.36	0.54	0.39	0.34
	RSA	0.62	0.17	-0.07	0.22	**0.08**	-0.01	0.48	0.29	0.22
Feature	CKA	0.67	0.34	**0.25**	0.14	-0.05	0.40	0.50	0.38	0.33
	PARC	0.69	**0.67**	0.14	**0.31**	-0.10	**0.65**	0.62	0.67	**0.46**
	RSA	**0.72**	0.65	0.02	0.28	0.02	0.19	**0.72**	**0.67**	0.41

Table 13: Kendall Tau correlation between test accuracy after all possible single-source distillations and task similarity associated with the source models. Similar to Table 2.

[1]. Thus, we consider each source model separately and compute the metrics independent of other source models. This has the added benefit of reducing the number of metric computations required as we do not need to compute the task similarity for all possible combinations of n models from S

possible (i.e. $\binom{n}{S}$), which grows fast with S.

Assume $\mathbf{X} \in \mathbb{R}^{N \times d_X}$ and $\mathbf{Y} \in \mathbb{R}^{N \times d_Y}$, and that $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ for and $\mathbf{L}_{ij} = l(\mathbf{y}_i, \mathbf{y}_j)$ where k, and l are two (similarity) kernels as well as $\mathbf{x}_i, \mathbf{x}_j$ and $\mathbf{y}_i, \mathbf{y}_j$ are rows of $\mathbf{X}$ and $\mathbf{Y}$, respectively. Then we have that CKA is defined as

$$\rho_{\text{CKA}}(\mathbf{X}, \mathbf{Y}) \overset{\text{def}}{=} \frac{\text{HSIC}(\mathbf{K}, \mathbf{L})}{\sqrt{\text{HSIC}(\mathbf{K}, \mathbf{K})\text{HSIC}(\mathbf{L}, \mathbf{L})}},$$

where $\mathbf{K}, \mathbf{L} \in \mathbb{R}^{N \times N}$ and HSIC is the Hilbert-Schmidt Independence Criterion,

$$\text{HSIC}(\mathbf{K}, \mathbf{L}) \overset{\text{def}}{=} \frac{1}{(N-1)^2} \text{tr}\left(\mathbf{K}\mathbf{H}_N\mathbf{L}\mathbf{H}_N\right), \quad \text{with}$$
$$\mathbf{H}_N \overset{\text{def}}{=} \mathbf{I}_N - \frac{1}{N}\mathbf{1}\mathbf{1}^\intercal.$$

In particular, if both k and l are linear kernels, then

$$\rho_{\text{CKA}}(\mathbf{X}, \mathbf{Y}) = \frac{\|\mathbf{Y}^\intercal\mathbf{X}\|_F^2}{\|\mathbf{X}^\intercal\mathbf{X}\|_F\|\mathbf{Y}^\intercal\mathbf{Y}\|_F},$$

where $\|\cdot\|_F$ is the Frobenius norm. We use the linear kernel throughout this paper and refer to Cortes et al. [3] for additional details on CKA.

For RSA, we consider the dissimilarity matrices given by

$$\mathbf{K}_{ij} \overset{\text{def}}{=} 1 - \text{pearson}(\mathbf{x}_i, \mathbf{x}_j) \quad \text{and}$$
$$\mathbf{L}_{ij} \overset{\text{def}}{=} 1 - \text{pearson}(\mathbf{y}_i, \mathbf{y}_j),$$

where $\mathbf{X}$ and $\mathbf{Y}$ are assumed normalized to have mean 0 and variance 1. We then compute RSA as the Spearman correlation between the lower triangles of $\mathbf{K}$ and $\mathbf{L}$,

$$\rho_{\text{RSA}}(\mathbf{X}, \mathbf{Y}) \overset{\text{def}}{=} \text{spearman}\left(\{\mathbf{K}_{ij} \mid i < j\}, \{\mathbf{L}_{ij} \mid i < j\}\right).$$

For additional details on RSA, we refer the reader to Dwivedi and Roig [4]. While Bolya et al. [2] introduces PARC alongside a heuristic and feature reduction, the PARC metric is almost identical to RSA. However, RSA was introduced to compute similarities between two sets of representations, and PARC was aimed at computing similarities between a set of representations and a set of labels associated with the dataset. Thus, in our use of PARC, it merely differs from RSA in the lack of normalization of $\mathbf{Y}$, which is assumed to be one-hot encoded vectors of class labels from the probe dataset.

A.9. Compute Requirements and Scalability

For DISTILLNEAREST and DISTILLWEIGHTED to be feasible in practice, we need to ensure that the computational costs of training and inference for both methods are reasonable and that it scales well with the size of $\mathcal{S}$.

Inference We note, that while both DISTILLNEAREST and DISTILLWEIGHTED use additional classifier head(s) during training, these are discarded at inference time, and no additional compute overhead remains. Thus, memory and compute requirements at inference time are identical to those of the original target model, and thereby the equivalent target model trained supervised.

Training We can separate the training procedure into two phases; a) estimation of task similarity metrics, and b) training of the target model with DISTILLNEAREST or DISTILLWEIGHTED. The majority of compute is typically needed for b) as is expected for the training of neural networks. For a) we estimate the task similarity metric for a single model based on the small annotated probe set (we use 500 samples). The computation of the metric itself is dominated by the forward pass, which is a single forward pass on each of the 500 samples, thus corresponding to less compute than 4 batches of training, where we typically train for thousands of batches. We thus consider phase a) as negligible as it is also reusable across multiple experiments for the target task. Furthermore, for b) we use additional compute in two parts of the training; 1) training of an additional classifier head per source model used (thus 1 for DISTILLNEAREST and $|\mathcal{S}|$ for DISTILLWEIGHTED), and 2) obtaining pseudo-labels for the unlabeled data. We note that for 224×224 inputs, a MobileNetV3 uses 0.24 GFLOPs as default, and each additional classifier head requires an additional approx. 0.0013 GFLOPs (for 1000 classes). Thus, we can attach about 3000 classifier heads (and thereby 3000 source models) to a MobileNetV3 before we require the same GFLOPs as a ResNet-50. Regarding 2), the pseudo-labels are obtained by a single forward pass by each source model over the unlabeled data. For very large source models, this can potentially be expensive, but the pseudo-labels can be reused across multiple experiments for the same target task. However, the computational requirements for this step highly depend on the set of source models $\mathcal{S}$, and is thereby hard to quantify.

B. Experimental Details

In the following, we provide some experimental details.

B.1. Main Experiments

Unless otherwise mentioned, we use SGD with a learning rate of 0.01, weight decay of 0.0001, batch size of 128, and loss weighting of $\lambda = 0.8$. We initialize our target models with the ImageNet pre-trained weights available in torchvision (https://pytorch.org/vision/stable/models) and consider 28 fine-tuned models from Bolya et al. [2] publicly available at github.com/dbolya/parc as our set of source models. The source models consist of each of the architectures

(AlexNet, GoogLeNet, ResNet-18, and ResNet-50) trained on CIFAR-10, Caltech101, CUB200, NABird, Oxford Pets, Stanford Dogs, and VOC2007. Note, we always exclude any source model trained on the particular target task, thus effectively reducing the number of source models for some target tasks. For FixMatch we use a batch size of 128 (with a 1:1 ratio of labeled to unlabeled samples for each batch) and fix the confidence threshold at 0.95 and the temperature at 1. We keep the loss weighting between the supervised loss and the unlabeled FixMatch loss at $\lambda = 0.8$.

B.2. VTAB Experiments

For each VTAB experiment, we consider the full training set (as introduced in Zhai et al. [7]) as the unlabeled set, $\mathcal{D}_\tau^u$, and the VTAB-1K subset as the labeled set, $\mathcal{D}_\tau^l$. We use the Pytorch implementation from Jia et al. [5] available at `github.com/KMnP/vpt`.

We use SGD with a learning rate of 0.005, weight decay of 0.0001, batch size of 128 equally split in 64 labeled and unlabeled samples, and loss weighting of $\lambda = 0.9$. We train our models for 100 epochs, where we define one epoch as the number of steps required to traverse the set of unlabeled target data, $\mathcal{D}_\tau^u$ when using semi-supervised methods, or merely as the number of steps to traverse the labeled set, $\mathcal{D}_\tau^l$, for supervised transfer methods. We initialize our target models with the BiT-M ResNet-50x1 model fine-tuned on ILSVRC-2012 from BiT [6] publicly available at `github.com/google-research/big_transfer`.

We consider the 19 BiT-M ResNet-50x1 models fine-tuned on the VTAB-1K target tasks from Kolesnikov et al. [6] as the set of source models. We always exclude the source model associated with the target task from the set of source models, and thus effectively have 18 source models available for each target task in VTAB. We use the PARC metric on the source model features to compute the source weighting, but also only use the top-5 highest-ranked source models to reduce the computational costs of training. Furthermore, we use $p = 9$ for DISTILLWEIGHTED.

C. Domain gap between source tasks, targets tasks and ImageNet

As is evident from Figure 3 and Table 1, both DISTILLNEAREST and DISTILLWEIGHTED do not yield notable improvements on e.g. ChestX and ISIC, but yield significant improvements on e.g. CUB200 and Oxford Pets. Notably, for the latter target tasks there are semantically similar source tasks present in our set of source models, while this is not true for the former target tasks. Hence, as one would expect, the availability of a source model trained on source tasks similar to the target tasks is important for cross-domain distillation to work well, which is expected to be true for both DISTILLNEAREST and DISTILLWEIGHTED. Indeed, the task similarity metrics considered in this paper all aim at measuring alignment between tasks, and if the alignment between source and target tasks is small, we do not expect to gain much from distillation. This is affirmed by our experiments in *e.g.* Table 1.

C.1. A note on potential data overlap between source and target tasks

Whenever any type of transfer learning is applied, including using ImageNet initializations, we (often implicitly) assume that the model we transfer from has not been trained on any data from the target test set. Although this assumption is often satisfied in practice due to domain gaps between the source and target task, utilizing initializations trained on e.g. ImageNet can potentially violate the assumption. This is due to the fact that ImageNet and many other modern publicly available datasets are gathered from various public websites and overlaps between samples in different datasets might occur.

Thus, it is natural to question whether the observed improvements are due to methodological advances or information leakage between source and target tasks. To ensure our advancements are valid we carefully remove any source model associated with the target task from the set of source models, $\mathcal{S}$. However, information leakage might still appear if *e.g.* there are identical samples in the target dataset and the source dataset or ImageNet. Despite large overlaps being improbable, it has been shown that there e.g. is a minor overlap (of at least 43 samples) between the training set of ImageNet and the test set of CUB200 (see e.g. `https://gist.github.com/arunmallya/a6889f151483dcb348fa70523cb4f578`). However, since the test set of CUB200 consists of 5794 samples, the presence of such a minor overlap should not affect the true performance of a model much.

In our experiments, we consistently compare our target models (initialized with ImageNet weights) to either identically initialized target models or source models initialized with either ImageNet weights or with weights from a source task. Hence, any potential gain from information leakage between ImageNet and a target task would bias both our results and the baselines, and thereby not affect our overall results. Furthermore, while an overlap between a source and target task might unfairly benefit the performance of our methods compared to IN+TRANSFER and IN+FIXMATCH, such an overlap would likely benefit the fine-tuned source models even more making this baseline even harder to outperform (see e.g. Figure 5 and Table 1). Thus, our results should be at most as biased as the baselines.

References

[1] 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. 4

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

[3] 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. 5

[4] 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. 5

[5] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim. Visual Prompt Tuning. *arXiv*, 2022. URL http://arxiv.org/abs/2203.12119. 6

[6] 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. 6

[7] X. Zhai, J. Puigcerver, A. Kolesnikov, P. Ruyssen, C. Riquelme, M. Lucic, J. Djolonga, A. S. Pinto, M. Neumann, A. Dosovitskiy, L. Beyer, O. Bachem, M. Tschannen, M. Michalski, O. Bousquet, S. Gelly, and N. Houlsby. A Large-scale Study of Representation Learning with the Visual Task Adaptation Benchmark. *arXiv*, 2019. URL http://arxiv.org/abs/1910.04867. 1, 6