Adaptive Consistency Regularization for Semi-Supervised Transfer Learning

Abulikemu Abuduweili[1,2]*, Xingjian Li[1,3]*, Humphrey Shi[2]†, Cheng-Zhong Xu[3], Dejing Dou[1]†

[1]Big Data Lab, Baidu Research, [2]SHI Lab, University of Oregon,
[3]State Key Lab of IOTSC, Department of Computer Science, University of Macau
{v_abuduweili,lixingjian,doudejing}@baidu.com, hshi3@uoregon.edu, czxu@um.edu.mo

Abstract

While recent studies on semi-supervised learning have shown remarkable progress in leveraging both labeled and unlabeled data, most of them presume a basic setting of the model is randomly initialized. In this work, we consider semi-supervised learning and transfer learning jointly, leading to a more practical and competitive paradigm that can utilize both powerful pre-trained models from source domain as well as labeled/unlabeled data in the target domain. To better exploit the value of both pre-trained weights and unlabeled target examples, we introduce ***adaptive consistency regularization*** *that consists of two complementary components: Adaptive Knowledge Consistency (AKC) on the examples between the source and target model, and Adaptive Representation Consistency (ARC) on the target model between labeled and unlabeled examples. Examples involved in the consistency regularization are adaptively selected according to their potential contributions to the target task. We conduct extensive experiments on popular benchmarks including CIFAR-10, CUB-200, and MURA, by fine-tuning the ImageNet pre-trained ResNet-50 model. Results show that our proposed adaptive consistency regularization outperforms state-of-the-art semi-supervised learning techniques such as Pseudo Label, Mean Teacher, and FixMatch. Moreover, our algorithm is orthogonal to existing methods and thus able to gain additional improvements on top of MixMatch and FixMatch. Our code is available at https://github.com/Walleclipse/Semi-Supervised-Transfer-Learning-Paddle.*

1. Introduction

Deep neural networks have achieved great success in supervised learning tasks especially in computer vision [21, 15]. Yet, this heavily relies on a large amount of labeled data. As data annotation is usually expensive and time-consuming, Semi-Supervised Learning (SSL), which pursues the goal of effectively leveraging both labeled and unlabeled data, is widely studied. Recent state-of-the-art methods can be roughly summarized in three categories, which are consistency based regularization [22, 42], entropy minimization [12] and pseudo label [23].

While most works focus on the general setting that training a randomly initialized model from scratch, we consider a more realistic setting utilizing the powerful pre-trained model which is adequately fit on large-scale datasets for general purposes such as ImageNet [6] and Places365 [54]. These pre-trained models are empirically proven to have excellent transferability on various down-streaming tasks [49] and can significantly improve the generalization capacity of target tasks especially when the sample size is relatively small. Moreover, they are free to fetch and can be efficiently fine-tuned to adapt to new tasks. A recent study [55] points out that the benefit of semi-supervised learning sometimes may be marginal when fine-tuning a pre-trained model on the target dataset. However, the investigation of a systematic solution on DNN-based semi-supervised transfer learning has rarely been delved into.

In this work, we propose a semi-supervised transfer learning framework beyond a simple combination of these two kinds of algorithms. We extend the effective idea of consistency regularization in semi-supervised learning to adapt to inductive transfer learning, where the pre-trained weight learned by the source task is available. Specifically, our method is composed of two essential components: (1) Adaptive Knowledge Consistency (AKC) on the examples between the source and target model. We utilize target examples to transfer knowledge from the pre-trained model and help generalize the target model inspired by recent studies about knowledge distillation [51] and transfer learning [24]. To cope with the risk of negative transfer [43] caused by the discrepancy between the source and target task, we use the knowledge adaptive sample importance for proper cross-task knowledge consistency regularization. Intuitively, we are inclined to select examples lying in the

*Equal contributions and by alphabetical order. † Correspondence.

trusted region of the source model. (2) Adaptive Representation Consistency (ARC) on the target model between labeled and unlabeled examples. In transfer learning applications, labeled examples are often insufficient and thus they are prone to be projected onto an inappropriate representation with only the supervision of their labels. To tackle this problem we utilize ample unlabeled examples to adjust the representation produced by supervised learning to the real target domain. This is achieved by minimizing their Maximum Mean Discrepancy (MMD) distance. Furthermore, we adaptively decide the sample set used for restricting the representation distance. An intuitive explanation about the motivation of ARC is showed in supplementary A.

We evaluate our method on several semi-supervised transfer learning settings considering various typical scenarios. We use popular datasets CIFAR-10, CUB-200-2011, MIT Indoor 67, and MURA,covering domains including objects, animals, scenes and, radiographs.

Our main contributions can be summarized in the following points.

- To the best of our knowledge, we are the first to propose an advanced end-to-end semi-supervised transfer learning framework for deep neural networks. Considering incorporating inductive transfer learning, our research is closer to the actual problems in practice. Previous empirical study [55] provided observations and understandings by directly combining SSL with fine-tuning, but did not develop effective algorithms.
- We introduce adaptive consistency regularization to improve semi-supervised transfer learning by exploiting the characteristics of both semi-supervised learning and transfer learning, including cross-task knowledge distillation with adaptive sample importance named Adaptive Knowledge Consistency and representation adaptation for supervised learning using selected unlabeled data as the reference named Adaptive Representation Consistency.
- We conduct extensive experiments and show that the proposed adaptive consistency regularization is superior to classic semi-supervised learning algorithms such as Pseudo Label, Mean Teacher, and MixMatch on various semi-supervised transfer learning tasks. Furthermore, our method is shown orthogonal to existing methods and can obtain additional improvements even on top of MixMatch and FixMatch, which combine several state-of-the-art SSL techniques.

2. Related Work

2.1. Deep Transfer Learning

Previous research [34] proposed a comprehensive survey dividing transfer learning into three categories, which are inductive transfer learning, transductive transfer learning, and unsupervised transfer learning, according to the relationship between the source and target domain, and whether examples are labeled in either domain. In the deep learning community, most concerned transfer learning tasks include fine-tuning, domain adaptation, and few-shot learning. In this paper, we focus on fine-tuning as the main method, which belongs to inductive transfer learning according to [34].

Fine-tuning. Previous research pointed out that deep neural networks well-trained on large scale datasets for general purpose show great transferability on various downstream tasks [49]. Thus fine-tuning a pre-trained model to adapt new tasks has become a popular paradigm for many real world applications [18]. To further improve the effectiveness, some methods are investigated to improve the knowledge exploitation of the pre-trained model during fine-tuning, instead of merely treating it as a better starting point than random initialization. For example, [24] argued that the starting point should be used as the reference to regularize the learned weight. [51] demonstrated that knowledge distillation through attention map can be applied to different tasks and useful to enhance the performance of transfer learning. [26] proposed a channel level attention for knowledge distillation from the source to target task. Besides the idea of utilizing the pre-trained model, there are studies from other perspective, such as sample selection [10, 5, 31], dynamic fine-tuning path selection [53, 14] and suppressing negative transfer [46, 25].

Domain Adaption. Different from fine-tuning, domain adaptation [38] copes with the problem of sample selection bias between the training and test data. An important concept in classic domain adaptation methods is to generate domain invariant representation over the training set. Some earlier studies [11, 17] proposed sample re-weighting algorithms to adjust the decision boundary learned by the training examples to adapt to the target domain. Another useful idea is to explicitly minimize the distribution distance between the source and target domain. This kind of methods [33, 28, 47] intend to learn a proper feature transformation that can simultaneously project both domains into a shared representation space. Our work is highly inspired by the critical ideas developed for domain adaptation such as sample re-weighting and representation adaptation, while the task is rather different.

Few-shot Learning. Few-shot learning has been paid to increasing attention in recent years as it aims at imitating human intelligence by which knowledge can be generalized provided only several examples. The mainstream research direction is related to meta learning [7, 40]. It is quite different from regular transfer learning paradigms that the transferred knowledge is how to learn rather than what (e.g. model parameter) has learned. Recent work [50] designed a semi-supervised few-shot learning framework TransMatch

by incorporating Imprinting and MixMatch. They demonstrated that utilizing unlabeled examples makes their framework surpass the purely supervised few-shot learning competitors.

2.2. Semi-Supervised Learning

There exist a vast number of classic works on semi-supervised learning, and most of them fall into one of the three main mechanisms[32]: consistency based regularization, entropy minimization, and pseudo label. All these methods share an intuition to use additional unlabeled data to exploit the underlying structure, which usually could hint the separation of samples whose labels we want to distinguish. We only briefly discuss the branch of consistency based regularization, which is the most related to our work.

Consistency regularization is based on the hypothesis that the decision boundary is not likely to pass through high-density areas. This hypothesis results in a specific principle that a sample and its close neighbours are expected to have the same label. This forms the basic motivation of consistency based methods, as well as many self-supervised learning approaches, which all care about the utilization of unlabeled data. For example, the Π model [22] arguments the input sample with different noises, and adds a regularization term to reduce the discrepancy between outputs with respect to the original input and its perturbed peers. Temporal Ensembling [22] and Mean Teacher [42] involve ensemble learning to promote the quality of labels of the perturbed samples. Specifically, they use the moving average weights or predictions. Recently, Interpolation Consistency Training (ICT) [44] improved the perturbation method by using Mixup with another unlabeled sample instead of adding random noise. This is regarded as a more efficient transformation when dealing with low-margin unlabeled points. MixMatch [2] further proposed artificial label sharpening for unlabeled data and mixing both labeled and unlabeled data in Mixup. FixMatch [41] continued the trend to combine diverse mechanisms for exploiting unlabeled examples.

Our work does not pursue to search for the best choice among those general semi-supervised learning algorithms in the transfer learning setting. Instead, we intend to develop more targeted strategies utilizing the properties of the combination of semi-supervised and transfer learning problems.

2.3. Semi-Supervised Transfer Learning

Semi-supervised transfer learning can be regarded as a natural extension of regular semi-supervised learning by taking a related auxiliary task into consideration or as an extension of regular transfer learning with only a proportion of the labeled target examples. There are few works targeting this sort of problem. Early work [39] investigated this problem under the setting of the traditional machine learning framework. They proposed an improved co-training method for inductive transfer learning with instance re-weighting according to the training error. Two diverse k-Nearest-Neighbour (kNN) learners with different values of k are trained collaboratively. Recently, [55] presented an empirical study showing that the gains from state-of-the-art SSL techniques decrease or sometimes even disappear compared with a fully-supervised baseline when we fine-tune the target task starting from a pre-trained model. While these observations pointed out the necessity of considering this more competitive and practice baseline, they did not aim at inventing a solution. [19] imposed the Lautum regularization with which they improved the pre-training stage using examples from both the source and target task. Although the accuracy outperforms several baselines, the requirements of accessing the source dataset and an extra pre-training for every target task are usually unrealistic.

Some recent studies investigated semi-supervised transfer learning on specific tasks. [48] discussed the task of rain removal with a framework of semi-supervised transfer learning. [47] introduced a semi-supervised domain adaptation method for semantic segmentation. [8] studied pseudo-labeling method on unsupervised domain adaptation for person re-identification.

Different from those works, this paper introduces a novel framework for general semi-supervised transfer learning.

3. The Proposed Framework

The flowchart of the proposed semi-supervised transfer learning is illustrated in Figure 1. Please check the more detailed illustration of the proposed framework in supplementary A.

Problem definition: In inductive transfer learning, we have the source dataset $\mathcal{D}_s$ and the target dataset $\mathcal{D}_t$ corresponding to different tasks. A typical deep neural network f can be split into two parts: a representation function F_θ and a task-specific function G_ϕ. F_θ is able to contain general knowledge if trained over a dataset with diverse semantics and thus is transferable. While G_ϕ has the particular architecture with respect to the task attribute such as the number of classes. We denote the parameters of the representation function (called feature extractor in our task) and task-specific function (called classifier in our task) pre-trained over the source dataset as θ^0 and ϕ^0 respectively. For the target dataset, we denote $\mathcal{D}_t^l = \{\mathbf{x}_l^{1,2\cdots n}\}$ as the labeled examples and $\mathcal{D}_t^u = \{\mathbf{x}_u^{1,2\cdots,m}\}$ as the unlabeled examples. Here we ignore the subscript s or t for a specific example $\mathbf{x}$ as we will only use the target dataset after the pre-training stage. We then define the complete target dataset $\mathcal{D}_t = \mathcal{D}_t^l \cup \mathcal{D}_t^u$ and its size is $n+m$. To solve the target task, we formalize the general form of the optimizing

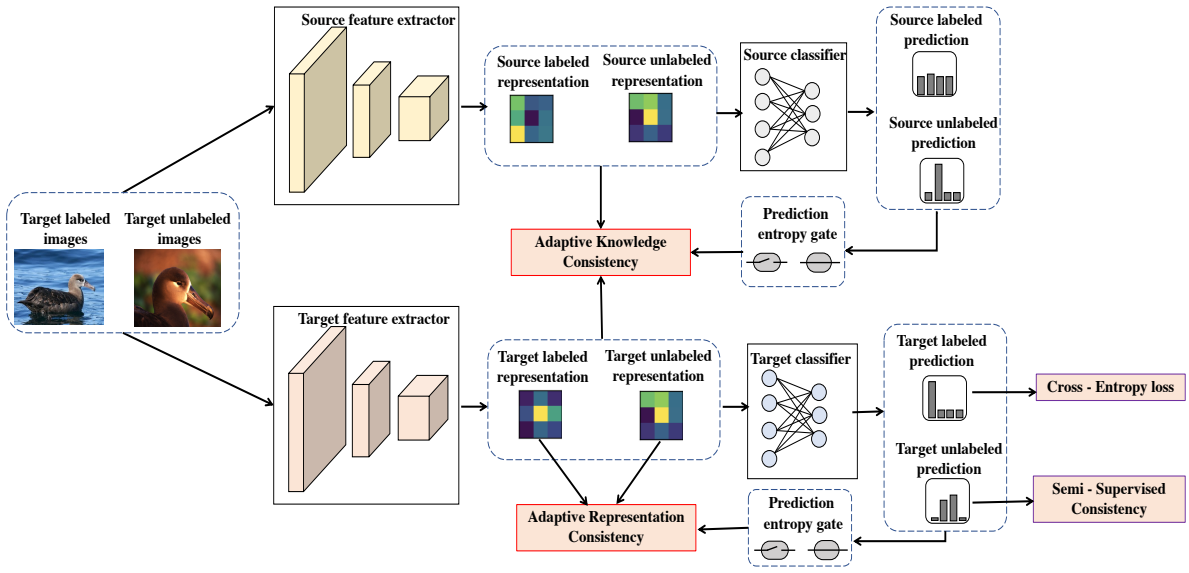

Figure 1. The framework of adaptive consistency regularization for semi-supervised and transfer learning.

objective as

$$\theta^*, \phi^* = \arg\min_{\theta,\phi} \sum_{i=1}^{n} L_{\mathrm{CE}}(\theta, \phi; \mathbf{x}_l^i) + R(\theta) \qquad (1)$$

, where L_{CE} is the commonly used cross-entropy loss indicating the prediction error and R refers to additional regularization related to the pre-trained parameter θ^0, ϕ^0 and the target dataset $\mathcal{D}_t$. Note that since a labeled example can be regarded as unlabeled if we ignore its label, we actually use $\mathcal{D}_t$ when we need a set of unlabeled examples.

3.1. Pre-training and Imprinting

We adopt a popular strategy to implement inductive transfer learning, which is to sequentially learn from the source and the target dataset. The first step is pre-training. The representation parameter of the target model is initialized with θ^0. We do not discuss other paradigms of utilizing the source dataset in this paper such as co-training the source and target dataset like [10].

Although the task-specific function G can not be shared directly, we borrow the idea of Imprinting from recent low-shot learning research [35]. Imprinting performs an informative initialization on G instead of random initialization. Such knowledge derived from the feature extractor F of the source model provides a much better starting point to the target model with immediate good classification performance.

3.2. Adaptive Knowledge Consistency

Knowledge distillation is widely studied with the original motivation of compressing complex ensembling models [16]. While recent studies reveal that knowledge distillation can also help improve the identical model [9] over the same task and even generalize a different task [51, 27, 26]. We adopt the method to distill the knowledge of the source model through the representation rather than the task-specific logits output, as the latter is not suitable for handling different tasks. While different from previous studies, we employ both labeled and unlabeled data as the bridge of knowledge transfer and impose adaptive sample importance to prevent negative transfer cause by the discrepancy between the two datasets. Specifically, we constrain the weighted Kullback–Leibler divergence (or mean square error) of outputs between the pre-trained feature extractor F_{θ^0} and the target feature extractor F_θ using the entire target dataset $\mathcal{D}_t$. In our setting, we denote $\mathcal{L} = \{\mathbf{x}_l^i\}^{B_l} \subset \mathcal{D}_t^l$ as a mini-batch of B_l labeled examples, and $\mathcal{U} = \{\mathbf{x}_u^i\}^{B_u} \subset \mathcal{D}_t^u$ as a mini-batch of B_u unlabeled examples. Formally, the regularization term of a mini-batch can be written as

$$R_{\mathrm{K}} = \frac{1}{B_l + B_u} \sum_{\mathbf{x}^i \in \mathcal{L} \cup \mathcal{U}} w_{\mathrm{K}}^i \, \mathrm{KL}(F_{\theta^0}(\mathbf{x}^i), F_\theta(\mathbf{x}^i)) \qquad (2)$$

To calculate the sample importance w_{K}^i, we leverage the pre-trained source model with the parameter θ^0 and ϕ^0. In detail, a target example $\mathbf{x}^i$ is fed forward the pre-trained model and we obtain the final output post-processed by the softmax operation, marked as $\mathbf{p}_s^i = G_{\phi^0}(F_{\theta^0}(\mathbf{x}^i))$. $\mathbf{p}_s^i$ is a 1-dimensional vector with the length equal to the number of source classes C_s. We get the weight of sample $\mathbf{x}^i$ by calculating the entropy of $\mathbf{p}_s^i$ as:

$$w_{\mathrm{K}}^i = \mathcal{G}(\mathrm{H}(\mathbf{p}_s^i)) = \mathcal{G}(-\sum_{j=1}^{C_s} \mathbf{p}_{s,j}^i \log(\mathbf{p}_{s,j}^i)). \qquad (3)$$

Where $\mathcal{G}$ is an entropy-gate function, which projects calculated entropy to a value of sample importance. Intuitively,

the entropy of the output as a probability on different classes indicates the confidence of the recognition with respect to the input. In other words, higher output confidence implies that the input sample is more likely to fall into the source model's trust region and consequently the knowledge about this sample is reliable to the target model. In our implementation, we perform a hard filter according to the sample importance with a pre-determined threshold value ϵ_{K} so as to reduce the extra computation burden. Sample importance w_{K}^i can be written as a binary value of:

$$w_{\mathrm{K}}^i = \mathrm{I}(\mathrm{H}(\mathbf{p}_s^i) \leq \epsilon_{\mathrm{K}}) \tag{4}$$

The sample importance $w_{\mathrm{K}}^i = 1$, only if the corresponding entropy is lower than pre-determined threshold $\mathrm{H}(\mathbf{p}_s^i) \leq \epsilon_{\mathrm{K}}$.

3.3. Adaptive Representation Consistency

In this part, we introduce another imposed regularizer named adaptive representation (distribution) consistency, by which we intend to tackle the problem of over-fitting the insufficient labeled target samples. Motivated by the fact that unlabeled samples themselves contain potential information about the data structure, we utilize unlabeled target samples to help labeled samples learn representations with stronger generalization ability. Different from knowledge distillation incorporating the alignment at the sample level, the representation consistency affects training at the distribution level. Specifically, we use the classical metric Maximum Mean Discrepancies (MMD) [3] to measure the distance between the representations of labeled and unlabeled data. Denoting $\mathbf{V} = \{\mathbf{v}^1, \mathbf{v}^2, ..., \mathbf{v}^n\}$ and $\mathbf{U} = \{\mathbf{u}^1, \mathbf{u}^2, ..., \mathbf{u}^m\}$ as random variable sets with distributions Q_v and Q_u, an unbiased estimate of the MMD between Q_v and Q_u compares the square distance between the empirical kernel mean embeddings as

$$\mathbf{MMD}(Q_v, Q_u) = \| \frac{1}{m} \sum_{i=1}^{m} \kappa(\mathbf{v}^i) - \frac{1}{n} \sum_{j=1}^{n} \kappa(\mathbf{u}^j) \|^2, \tag{5}$$

where κ refers to the kernel, as which a Gaussian radial basis function (RBF) is usually used in practice [13, 29].

In our case, we need to measure the MMD between labeled representation $\{F_\theta(\mathbf{x}_l^i) | \mathbf{x}_l^i \in \mathcal{L}\}$ distribution and unlabeled representation $\{F_\theta(\mathbf{x}_u^i) | \mathbf{x}_u^i \in \mathcal{U}\}$ distribution. Nevertheless, this restrain raises a severe risk because the target model is progressively learned. Thus even the representation distribution obtained by sufficient unlabeled examples is inaccurate at earlier stages of the training procedure. To overcome this kind of problem, we involve an adaptive sample selection method similar to that in adaptive knowledge consistency. Specifically, we compute the entropy of the softmax output given a sample as the input and regard the entropy as the target model's confidence on this sample. Only confident samples will be employed to regularize the representation of labeled data. In detail, a labeled example $\mathbf{x}_l^i$ (and an unlabeled example $\mathbf{x}_u^i$) is fed forward the target model and we obtain the final output as $\mathbf{p}_l^i = G_\phi(F_\theta(\mathbf{x}_l^i))$ (and $\mathbf{p}_u^i = G_\phi(F_\theta(\mathbf{x}_u^i))$), then we get the gate state (whether selection or not) of the example by calculating the entropy of prediction as $\mathbf{H}(\mathbf{p}_l^i)$ (and $\mathbf{H}(\mathbf{p}_u^i)$) considering predefined threshold value ϵ_{R}. Denoting set of selected labeled representation as $\mathcal{F}_l$ and set of selected unlabeled representation as $\mathcal{F}_u$:

$$\begin{aligned} \mathcal{F}_l &= \{F_\theta(\mathbf{x}_l^i) | x_l^i \in \mathcal{L} \text{ and } \mathbf{H}(\mathbf{p}_l^i) \leq \epsilon_{\mathrm{R}}\} \\ \mathcal{F}_u &= \{F_\theta(\mathbf{x}_u^i) | x_u^i \in \mathcal{U} \text{ and } \mathbf{H}(\mathbf{p}_u^i) \leq \epsilon_{\mathrm{R}}\} \end{aligned} \tag{6}$$

Note that the sample selection result is adaptively changing as the target model progressively fits more training examples. Considering that the number of selected samples in a mini-batch may not be adequate to calculate a convinced distribution, we impose a replay buffer to save recent selected confident examples. The replay buffer enables us to calculate MMD with more data, and which is helpful to approximate full representation distribution with recent some mini-batches representation distribution. The pseudo-code of the replay buffer is quite straightforward, as following:

$$\begin{aligned} &\mathrm{Labeled\_Buffer.update}(\mathcal{F}_l) \\ &\mathrm{Unlabeled\_Buffer.update}(\mathcal{F}_u) \\ &\mathcal{F}_l^\star = \mathrm{Labeled\_Buffer.get\_last\_k}() \\ &\mathcal{F}_u^\star = \mathrm{Unlabeled\_Buffer.get\_last\_k}() \end{aligned} \tag{7}$$

Denoting $Q_{\mathcal{F}_l^\star}$ and $Q_{\mathcal{F}_u^\star}$ as the representation distribution generated from $\mathcal{F}_l^\star$ and $\mathcal{F}_u^\star$, we give the adaptive representation consistency as the following form:

$$R_{\mathbf{R}} = \mathbf{MMD}(Q_{\mathcal{F}_l^\star}, Q_{\mathcal{F}_u^\star}). \tag{8}$$

3.4. Summarization of the Framework

We finally present the complete adaptive consistency regularization consisting of AKC and ARC as

$$R(\theta) = \lambda_{\mathbf{K}} R_{\mathbf{K}} + \lambda_{\mathbf{R}} R_{\mathbf{R}}. \tag{9}$$

Where $\lambda_{\mathbf{K}}$ and $\lambda_{\mathbf{R}}$ are weighted factors for AKC and ARC. If we incorporate cross-entropy loss L_{CE} for labeled data and semi-supervised consistency loss $L_{\mathbf{S}}$ for unlabeled data (just like MixMatch, FixMatch, Pseudo-labeling ...), then the final loss function would become:

$$\begin{aligned} L(\theta, \phi) = &\frac{1}{n} \sum_{i=1}^{n} L_{\mathrm{CE}}(\theta, \phi; \mathbf{x}_l^i) + \lambda_{\mathbf{S}} L_{\mathbf{S}}(\{\mathbf{x}_u^i\}) + \\ &\lambda_{\mathbf{K}} R_{\mathbf{K}}(\{\mathbf{x}_l^i\}, \{\mathbf{x}_u^i\}) + \lambda_{\mathbf{R}} R_{\mathbf{R}}(\{\mathbf{x}_l^i\}, \{\mathbf{x}_u^i\}) \end{aligned} \tag{10}$$

Where $\lambda_{\mathbf{S}}$ is a weighted factor for semi-supervised consistency loss. After initializing with the pre-trained source model and imprinting, the remaining fine-tuning is performed in an end-to-end manner.

4. Experiments

4.1. Experimental setup

4.1.1 Dataset configuration

We evaluate our proposed adaptive consistency regularization methods and compare with state-of-the-art semi-supervised learning methods on several public datasets including the commonly used semi-supervised learning dataset CIFAR-10 [20] and transfer learning benchmarks CUB-200-2011[45], MIT Indoor-67[36] and musculoskeletal radiographs dataset MURA[37]. ImageNet[6] is used as the source task. Note that CIFAR-10, Indoor-67 and CUB-200-2011 have some classes semantically overlaps with ImageNet, while MURA is a medical image dataset with a large domain mismatch from ImageNet. Detailed descriptions about these datasets are listed in supplementary B.1.

4.1.2 Baseline

We compare proposed adaptive consistency regularization methods with the following state-of-the-art semi-supervised learning methods. In order to make a fair comparison in semi-supervised transfer learning tasks, we incorporate these semi-supervised learning methods with the same strategies including initialization with imprinting and fine-tuning all layers.

- Standard fine-tuning on labeled dataset: This is equivalent to a pure supervised manner where unlabeled examples are not used.
- Pseudo-labeling [23]: It proceeds by producing "pseudo-labels" for unlabeled training set using the prediction function itself over the course of training.
- Mean-teacher [42]: It obtains more stable target predictions for unlabeled training set. Specifically, it sets the target labels using an exponential moving average of parameters from previous training steps. The representation consistency between the original and perturbed unlabeled samples is encouraged, as well as the standard cross-entropy minimization for labeled samples.
- MixMatch [2]: In addition to the consistency regularization, it proposes artificial label sharpening for pseudo-labeling on unlabeled data and mixing both labeled and unlabeled data in Mixup during the process of fine-tuning.
- FixMatch [41]: FixMatch further improves on top of the above techniques. It computes an artificial label given a weakly augmented version of a given unlabeled image. Then it uses the pseudo-label to enforce the cross-entropy loss against the model's output for a strongly-augmented version of the unlabeled image.

It should be noted that our proposed adaptive consistency regularization techniques are theoretically compatible with other semi-supervised methods. Thus, we also evaluate our proposed regularization techniques integrated with MixMatch or FixMatch.

4.1.3 Training strategy

On the transfer learning benchmarks, we use ImageNet as our source dataset and use ResNet-50[15] pre-trained model as our source model by default unless explicitly specified. We fine-tune the ImageNet pre-trained model on CUB-200-2011, Indoor-67, and MURA datasets with labeled and unlabeled samples. We use SGD with momentum as the optimizer to train the target model 200 epochs. The momentum rate is set to be 0.9, the initial learning rate is 0.001 (except that the initial learning rate is 0.01 for CUB-200-2011) and the mini-batch size is 64 for both labeled and unlabeled dataset. For a learning rate schedule, we use a cosine learning rate decay[30] which sets the learning rate to

$$\eta_t = \eta_0 \cos(\frac{7\pi t}{16T}) \tag{11}$$

where η_0 is the initial learning rate, t is the current training step, and T is the total number of training steps. For our semi-supervised fine-tuning method, we set the parameters of AKC and ARC as follows. We set the regularization weight factors as $\lambda_{\mathbf{K}} = 1$ and $\lambda_{\mathbf{R}} = 30$, and adaptive thresholds as $\epsilon_{\mathbf{K}} = 0.7 \cdot \log(C_s)$ and $\epsilon_{\mathbf{R}} = 0.7 \cdot \log(C_t)$. Where C_s and C_t refer to the class number of source dataset and target dataset.

On the CIFAR-10 experiment, following the experiment setting by [41], we use the same network architecture Wide ResNet-28-2 [52] and training protocol, including the optimizer, learning rate schedule, data preprocessing, across all SSL methods. In the pre-training procedure, we train our Wide ResNet-28-2 model on ImageNet downsampled to 32×32 [4] (the native image size of CIFAR-10). The top-1 classification error rate is reported for clear demonstration.

4.2. Results

4.2.1 Results on CUB-200-2011

The results of adaptive knowledge consistency (AKC), adaptive representation consistency (ARC), and baseline methods on CUB-200-2011 dataset are listed in Table 1. The method of combining AKC with ARC achieved best or comparable performance among previous-best baseline methods, especially in the case that labeled samples are fewer. For example, when the size of the labeled dataset is 200, the AKC+ARC method relatively improves the accuracy by 27.8% compared to MixMatch. One of the advantages of our proposed method is its compatibility. AKC and

Methods \#label	2000	1000	400	200
Supervised labeled	68.29	53.26	28.82	17.90
Pseudo label	71.38	49.50	25.65	10.42
Mean teacher	70.19	51.78	27.01	13.79
MixMatch	73.84	60.56	32.79	22.66
FixMatch	72.76	58.30	31.03	21.86
AKC	71.33	58.42	38.71	28.57
ARC	72.95	61.01	41.13	28.47
AKC+ARC	73.65	62.01	41.69	28.96
MixMatch+AKC+ARC	77.51	67.26	43.80	29.55
FixMatch+AKC+ARC	75.59	63.36	40.83	28.25

Table 1. Classification accuracy of proposed AKC, ARC, and baselines on CUB-200-2011 dataset.

Methods \#label	1000	400
Supervised labeled	71.95	67.54
Pseudo label	73.99	67.56
Mean teacher	72.20	65.53
MixMatch	73.85	68.94
FixMatch	75.10	69.43
AKC	73.78	70.44
ARC	73.91	71.19
AKC+ARC	73.94	71.34
MixMatch +AKC+ARC	74.72	70.94
FixMatch +AKC+ARC	76.60	72.14

Table 2. Classification accuracy of proposed AKC, ARC, and baselines on MURA dataset.

Method \#label	4000	250	40
Supervised labeled	7.85	15.92	27.75
Pseudo label	7.04	12.92	25.62
Mean teacher	6.43	14.03	24.67
MixMatch	5.52	10.01	21.50
FixMatch	4.24	5.04	9.05
AKC	6.72	14.49	24.51
ARC	7.07	15.19	25.13
AKC+ARC	6.55	13.93	24.17
MixMatch +AKC+ARC	4.92	8.95	18.90
FixMatch +AKC+ARC	4.19	4.99	7.62

Table 3. Comparison of error rate using proposed AKC, ARC, and baselines on CIFAR-10 dataset.

ARC regularization terms could be combined with other semi-supervised learning methods, like MixMatch and FixMatch. By utilizing AKC and ARC regularization techniques in MixMatch, the performance increased notably. For the fine-tuning with 2000 (and 200) labeled sample, the performance of MixMatch is increased by 5.0% (and 30.40%) than vanilla MixMatch. We speculate that one major reason for the effectiveness of AKC and ARC is that AKC and ARC could effectively prevent severe over-fitting when the number of labeled examples is small. *

Results on Indoor-67 are presented in supplementary B.2 .

4.2.2 Results on MURA

The results of MURA dataset are listed in Table 2. Although MURA is a medical image dataset with a large domain mismatch from ImageNet, the AKC and ARC can also improve the performance. By utilizing AKC and ARC regularization techniques in FixMatch, the method of FixMatch+AKC+ARC achieves the best performance in both cases of 1000 and 400 labeled samples.

4.2.3 Results on CIFAR-10

The results of adaptive knowledge consistency (AKC), adaptive representation consistency (ARC), and baseline methods on CIFAR-10 dataset are listed in Table 3. By utilizing AKC and ARC regularization techniques in FixMatch, the method of FixMatch+AKC+ARC achieves the best performance in both cases of 4000, 250, and 40 labeled samples. By utilizing AKC and ARC regularization techniques in MixMatch, the performance increases notably. When fine-tuning with 250 labeled samples, the error rate of MixMatch is decreased by 10.59% if we impose AKC and ARC in it. For the previous-best method FixMatch, the proposed method still improves the performance, especially in very few labeled data training.

Note that when 4000 examples (only 8% of labeled data) are labeled, FixMatch achieves even lower top-1 error rate (4.24%) than fully supervised learning from scratch using all 50000 examples (5.01%), indicating that FixMatch employs advanced techniques beyond the mere utilization of unlabeled data. Therefore, it's reasonable that additional improvements will not be remarkable on top of such a competitive baseline. We presented the effectiveness of transfer learning in low-data semi-supervised learning on CIFAR-10 in supplementary B.5.

4.2.4 Ablation Study

Adaptiveness of our proposed AKC and ARC regularization methods is affected by threshold value ϵ_{K} and ϵ_{R}. If $\epsilon_{\mathrm{K}} = 0$ (and $\epsilon_{\mathrm{R}} = 0$), the AKC (and ARC) equals to being removed since non of the sample was selected to calculate the regularization term. If $\epsilon_{\mathrm{K}} = \max(\mathrm{H}(\mathrm{p_s})) = \log(C_s)$ (and $\epsilon_{\mathrm{R}} = \max(\mathrm{H}(\mathrm{p_t})) = \log(C_t)$), the AKC (and ARC) degenerates to non-adaptive regularization terms with calculating consistency on all samples.

We investigate the performance of AKC under differ-

*We notice that FixMatch is not superior to MixMatch on CUB-200-2011. This observation is partially consistent with the empirical investigation that the benefit of SSL algorithms may be marginal when we transfer the source model to a similar target task [55].

$\epsilon_{\mathrm{K}}/\log(C_s)$	0	0.3	0.5	0.7	1.0
2000 labels	68.29	70.33	70.74	71.33	70.70
400 labels	28.82	31.27	33.51	38.71	34.62

Table 4. Performance of proposed AKC under different ϵ_{K} on CUB-200-2011 dataset with 2000 and 400 labeled samples.

ent ϵ_{K} on CUB-200-2011 dataset, as shown in Table 4. As can be seen, AKC achieves better performance with $\epsilon_{\mathrm{K}} = 0.7 \cdot \log(C_s)$. This shows the effectiveness of "adaptive" method , especially for the case of 400 labeled sample, adaptive knowledge consistency (with $\epsilon_{\mathrm{K}} = 0.7 \cdot \log(C_s)$) outperformed standard "non-adaptive" knowledge consistency (with $\epsilon_{\mathrm{K}} = \log(C_s)$) by 11.8%.

We also investigate the performance of ARC under different ϵ_{R} on CUB-200-2011 dataset and get similar observations as AKC, as shown in Table 5. Thanks to adaptiveness, adaptive representation consistency performs better than non-adaptive representation consistency which uses all samples. In the case of 400 labeled sample, ARC with $\epsilon_{\mathrm{R}} = 0.5 \cdot \log(C_t)$ outperforms non-adaptive representation consistency by 5.7%.

$\epsilon_{\mathrm{R}}/\log(C_t)$	0	0.3	0.5	0.7	1.0
2000 labels	68.29	69.57	71.73	72.95	71.77
400 labels	28.82	34.01	41.88	41.13	39.63

Table 5. Performance of proposed ARC under different ϵ_{R} on CUB-200-2011 dataset with 2000 and 400 labeled samples.

The actual sample selected ratio in ARC and AKC is shown in Figure 2 on CUB-200-2011 dataset experiment with 2000 labeled samples. As can be seen, the sample selected ratio for ARC is gradually increasing in the first 10 epochs from 0.3 to 0.9. Which can be regarded as a kind of curriculum learning[1]. In the earlier stage of training, only a few high confident samples were used for labeled and unlabeled distribution consistency regularization. After 10 epochs of training, the sample ratio converges to near 0.9, indicating that some of the low-confident samples are never used for ARC regularization. This process would be beneficial for training since some "very hard" or abnormal samples might be harmful for generalization. The sample selected rate of AKC is stable during training as the source model is frozen during fine-tuning.

Additional experiments on increased accuracy after utilizing AKC and ARC are presented in supplementary B.3.

4.3. Beyond semi-supervised transfer learning

Albeit the proposed Adaptive Knowledge Consistency (AKC) and Adaptive Representation Consistency (ARC) regularization methods are targeted at the semi-supervised transfer learning scenario, the application of those two regularization methods is not merely limited to semi-supervised transfer learning tasks.

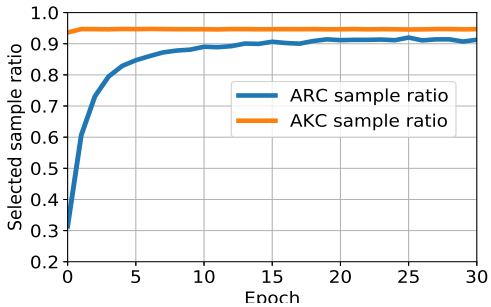

Figure 2. Effective sample ratio used in calculating ARC and AKC.

Method	Standard	AKC	ARC	ARC+AKC
Accuracy	81.77	82.79	82.54	83.52

Table 6. Results of AKC and ARC on CUB-200-2011 supervised transfer learning.

The AKC regularization can be incorporated with other supervised or unsupervised transfer learning methods since it does not require any label of the target data. Additionally, it is also suitable for tasks which involves multiple models, such as knowledge distillation from a big teacher model to a small student model.

The ARC regularization is also applicable for semi-supervised learning tasks training from scratch. Additionally, it can also be used in fully supervised learning, where we can easily regard the labeled set as the unlabeled set. Table 6 shows the result of the AKC and ARC regularization methods in fully supervised transfer learning in CUB-200-2011 dataset. Both AKC and ARC improve the performance of standard transfer learning.

5. Conclusion

In this paper, we propose two regularization methods: Adaptive Knowledge Consistency (AKC) between the source and target model and Adaptive Representation Consistency (ARC) between labeled and unlabeled examples. We show that AKC and ARC are competitive among state-of-the-art SSL methods. Furthermore, by incorporating AKC and ARC with other SSL methods, we achieve the best performance among several baseline methods on various transfer learning benchmarks. Additionally, our adaptive consistency regularization methods could be used for more general transfer learning and (semi-) supervised learning frameworks.

6. Acknowledgements

This work is supported in part by the Science and Technology Development Fund of Macau SAR (File no. 0015/2019/AKP to Chengzhong Xu).

References

[1] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009. 8

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 5049–5059, 2019. 3, 6

[3] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006. 5

[4] Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*, 2017. 6

[5] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4109–4118, 2018. 2

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 6

[7] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017. 2

[8] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6112–6121, 2019. 3

[9] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International Conference on Machine Learning*, pages 1607–1616. PMLR, 2018. 4

[10] Weifeng Ge and Yizhou Yu. Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 10–19, 2017. 2, 4

[11] Boqing Gong, Kristen Grauman, and Fei Sha. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *Proceedings of the International Conference on Machine Learning*, pages 222–230, 2013. 2

[12] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Proceedings of the Advances in neural information processing systems*, pages 529–536, 2005. 1

[13] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012. 5

[14] Yunhui Guo, Honghui Shi, Abhishek Kumar, Kristen Grauman, Tajana Rosing, and Rogerio Feris. Spottune: transfer learning through adaptive fine-tuning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4805–4814, 2019. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 6

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *stat*, 1050:9, 2015. 4

[17] Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex J Smola. Correcting sample selection bias by unlabeled data. In *Proceedings of the Advances in neural information processing systems*, pages 601–608, 2007. 2

[18] Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016. 2

[19] Daniel Jakubovitz, Miguel RD Rodrigues, and Raja Giryes. Lautum regularization for semi-supervised transfer learning. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3

[20] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the Advances in Neural Information Processing Systems*, 2012. 1

[22] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 1, 3

[23] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Proceedings of the International conference on machine learning, Workshop on challenges in representation learning*, volume 3, 2013. 1, 6

[24] Xuhong Li, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *35th International Conference on Machine Learning*, 2018. 1, 2

[25] Xingjian Li, Haoyi Xiong, Haozhe An, Cheng-Zhong Xu, and Dejing Dou. Rifle: Backpropagation in depth for deep transfer learning through re-initializing the fully-connected layer. In *International Conference on Machine Learning*, pages 6010–6019. PMLR, 2020. 2

[26] Xingjian Li, Haoyi Xiong, Hanchao Wang, Yuxuan Rao, Liping Liu, and Jun Huan. Delta: Deep learning transfer using feature map with attention for convolutional networks. In *International Conference on Learning Representations*, 2018. 2, 4

[27] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 4

[28] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the International conference on machine learning*, pages 97–105. PMLR, 2015. 2

[29] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2208–2217. JMLR. org, 2017. 5

[30] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *Learning*, 10:3, 2016. 6

[31] Jiquan Ngiam, Daiyi Peng, Vijay Vasudevan, Simon Kornblith, Quoc V Le, and Ruoming Pang. Domain adaptive transfer learning with specialist models. *arXiv preprint arXiv:1811.07056*, 2018. 2

[32] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Proceedings of the Advances in neural information processing systems*, pages 3235–3246, 2018. 3

[33] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2010. 2

[34] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009. 2

[35] Hang Qi, Matthew Brown, and David G Lowe. Low-shot learning with imprinted weights. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5822–5830, 2018. 4

[36] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 413–420. IEEE, 2009. 6

[37] Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L Ball, et al. Mura dataset: Towards radiologist-level abnormality detection in musculoskeletal radiographs. *Hand*, 1(602):2–215. 6

[38] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *Proceedings of the European conference on computer vision*, pages 213–226. Springer, 2010. 2

[39] Yuan Shi, Zhenzhong Lan, Wei Liu, and Wei Bi. Extending semi-supervised learning methods for inductive transfer learning. In *Proceedings of the Ninth IEEE international conference on data mining*, pages 483–492. IEEE, 2009. 3

[40] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Proceedings of the Advances in neural information processing systems*, pages 4077–4087, 2017. 2

[41] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 33, 2020. 3, 6

[42] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proceedings of the Advances in neural information processing systems*, pages 1195–1204, 2017. 1, 3, 6

[43] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010. 1

[44] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. *stat*, 1050:19, 2019. 3

[45] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 6

[46] Ruosi Wan, Haoyi Xiong, Xingjian Li, Zhanxing Zhu, and Jun Huan. Towards making deep transfer learning never hurt. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 578–587. IEEE, 2019. 2

[47] Zhonghao Wang, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-Mei Hwu, Thomas S Huang, and Honghui Shi. Alleviating semantic-level shift: A semi-supervised domain adaptation method for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 936–937, 2020. 2, 3

[48] Wei Wei, Deyu Meng, Qian Zhao, Zongben Xu, and Ying Wu. Semi-supervised transfer learning for image rain removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3877–3886, 2019. 3

[49] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Proceedings of the Advances in neural information processing systems*, pages 3320–3328, 2014. 1, 2

[50] Zhongjie Yu, Lin Chen, Zhongwei Cheng, and Jiebo Luo. Transmatch: A transfer-learning scheme for semi-supervised few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12856–12864, 2020. 2

[51] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016. 1, 2, 4

[52] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference 2016*. British Machine Vision Association, 2016. 6

[53] Yinghua Zhang, Yu Zhang, and Qiang Yang. Parameter transfer unit for deep neural networks. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 82–95. Springer, 2019. 2

[54] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 1

[55] Hong-Yu Zhou, Avital Oliver, Jianxin Wu, and Yefeng Zheng. When semi-supervised learning meets transfer learning: Training strategies, models and datasets. *arXiv preprint arXiv:1812.05313*, 2018. 1, 2, 3, 7