Semi-Supervised Learning by Augmented Distribution Alignment

Qin Wang[1], Wen Li[1,*], Luc Van Gool[1,2]

[1]ETH Zurich [2]KU Leuven

qwang@student.ethz.ch {liwen,vangool}@vision.ee.ethz.ch

Abstract

In this work, we propose a simple yet effective semi-supervised learning approach called Augmented Distribution Alignment. We reveal that an essential sampling bias exists in semi-supervised learning due to the limited number of labeled samples, which often leads to a considerable empirical distribution mismatch between labeled data and unlabeled data. To this end, we propose to align the empirical distributions of labeled and unlabeled data to alleviate the bias. On one hand, we adopt an adversarial training strategy to minimize the distribution distance between labeled and unlabeled data as inspired by domain adaptation works. On the other hand, to deal with the small sample size issue of labeled data, we also propose a simple interpolation strategy to generate pseudo training samples. Those two strategies can be easily implemented into existing deep neural networks. We demonstrate the effectiveness of our proposed approach on the benchmark SVHN and CIFAR10 datasets. Our code is available at https://github.com/qinenergy/adanet.

1. Introduction

Semi-Supervised Learning (SSL) aims to learn a robust model with a limited number of labeled samples and a abundant number of unlabeled samples. As a classical learning paradigm, it has gained many interests from both machine learning and computer vision communities. Many approaches have been proposed in recent decades, including label propagation, graph regularization, *etc*. [7, 6, 3, 22, 5, 55]. Recently, there is an increasing interest in training deep neural networks in the semi-supervised learning scenario[33, 47, 32, 37, 39, 11, 10]. This is partially due to the data-intensive nature of the conventional deep learning techniques, which often impose heavy demands on data annotation and bring high cost.
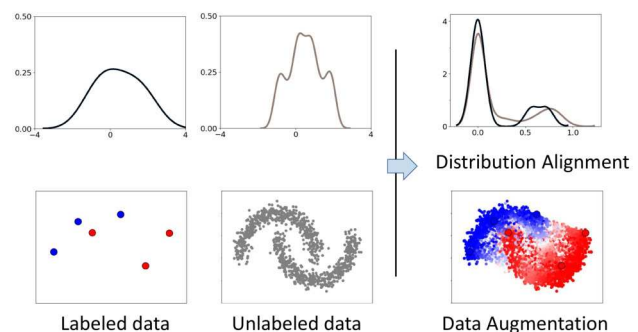

Figure 1. Illustration of the empirical distribution mismatch between labeled and unlabeled samples with the two-moon data. The labeled and unlabeled samples are shown in the **bottom left** and **bottom middle** figures, and the kernel density estimations of their x-axis projection are plotted in the **top left** and **top middle** figures, respectively. Our approach aims to address the empirical distribution mismatch by aligning sample distributions in the latent space (**top right**) and augmenting training samples with interpolation between labeled and unlabeled data (**bottom right**).

While many strategies have been proposed to utilize the unlabeled data for boosting the model performance, the essential sampling bias issue in SSL has rarely been discussed in the literature. That is, *the empirical distribution of labeled data often deviates from the true samples distribution*, due to the limited sampling size of labeled data. We illustrate this issue with the classical two-moon data in Figure 1, in which we plot 6 labeled samples (bottom left) and $1,000$ unlabeled samples (bottom middle). It can be observed the two-moon structure is well depicted by the unlabeled samples. However, due to the randomness in sampling and the small sample size, it can hardly tell the underlying distribution with the labeled data, though it is also sampled from the same two-moon distribution. In terms of empirical distribution, this also leads to a considerable difference between labeled and unlabeled data, as shown by the density estimation results on their x-axis projection (top left and top middle).

---

*The corresponding author

Similar empirical distribution mismatch is also observed in real world datasets for SSL (see Section 5.3). As observed in domain adaptation works, the model performance can often be significantly degraded when applying on a sample set with considerable empirical distribution difference. Therefore, the SSL models could also be potentially affected by the empirical distribution mismatch between labeled and unlabeled data when exploiting different SSL strategies, *e.g.*, label propagation from labeled data to unlabeled data.

To tackle this issue, we propose to explicitly reduce the empirical distribution mismatch in SSL. Specifically, we develop a simple yet effective approach called Augmented Distribution Alignment. On one hand, we adopt the adversarial training strategy to minimize the distribution distance between labeled and unlabeled data, such that the feature distributions are well aligned in the latent space, as illustrated in the top right of Figure 1. On the other hand, to alleviate the small sampling size issue and enhance the distribution alignment, we also propose a data augmentation strategy to generate pseudo samples by interpolating between labeled and unlabeled training sets, as illustrated in the bottom right of Figure 1. It is also worth mentioning that both strategies can be implemented easily, where the adversarial training could be achieved with a simple gradient reverse layer, and the data augmentation can be implemented by interpolation. Thus, they can be readily incorporated into existing neural networks for SSL with little effort. We demonstrate the effectiveness of our proposed approach on the benchmark SVHN and CIFAR10 datasets, on which we achieve new state-of-the-art classification performance.

Our contributions are summarized as follows:

- We offer a new perspective of empirical distribution mismatch to understand semi-supervised learning. The empirical distribution mismatch problem commonly exists in SSL scenarios, however, has not been revealed by existing semi-supervised learning works.

- We propose an augmented distribution alignment approach to explicitly address the empirical distribution mismatch for SSL.

- Our approach can be easily implemented into existing neural networks for SSL with little efforts.

- Despite of the simplicity, our proposed approach achieves new state-of-the-art classification performance on the the benchmark SVHN and CIFAR10 datasets for the SSL task.

2. Related Work

Semi-supervised learning: As a classical learning paradigm, many works have been proposed for semi-supervised learning with various methods, including label propagation, graph regularization, co-training, *etc*. [52, 42, 6, 36, 22, 5, 29, 1]. We refer interested readers to [55] for a comprehensive survey. Recently, there is an increasing interest in training deep neural networks in the semi-supervised learning scenario[47, 32, 37, 39, 11, 10]. This is partially due to the data-intensive nature of the conventional deep learning techniques, which often impose heavy demands on data annotation and bring high cost. Different models have been designed for deep semi-supervised learning. For example, [32, 47, 37] proposed to add small perturbations to unlabeled data, and enforce a consistency regularization [39] on the output of model. Other works [10, 11] adopt the idea of self-training and used propagated labels with a memory module or regularized by training speed. The ensemble approach was also explored, where [32] used an averaged prediction using the outputs of the network-in-training over time to regularize the model, while [47] instead used accumulated parameters to for prediction.

Different from above works, we tackle the SSL problem with a new perspective of empirical distribution mismatch, which was rarely discussed in the literature. By simply dealing with the distribution mismatch, we show that our newly proposed augmented distribution alignment with vanilla neural networks performs competitively with the state-of-the-arts SSL methods. Moreover, since we deal the SSL problem in a new way, our approach is potentially complementary to those approaches, and is shown to be able to further boost their performance.

Sampling bias problem: Sampling bias was usually discussed in the literature under the supervised learning and domain adaptation scenarios [44, 14, 27, 13]. Many works have been proposed to measure or address the sampling bias in the learning process [15, 16, 34, 45]. Recently, following the generative adversarial networks [20], the adversarial training strategy was widely used to address the empirical distribution mismatch in domain adaptation [16, 48, 51, 54, 9, 19, 8]. Although people generally assume samples in two domains are sampled from two different distributions, while in SSL the labeled and unlabeled samples are from the identical distribution, the techniques for reducing domain distribution mismatch used in domain adaptation can be readily used to solve the empirical distribution mismatch in SSL. In this work, we employ the adversarial training strategy proposed in [16]. A potential challenge as discussed in this paper is the small sample size of labeled data might lead to a lack of supports problem when aligning distribution, for which we additionally employ a sample augmentation strategy.

Other related works: Our work is also related to the recent proposed interpolation based data augmentation methods for training neural networks [53, 28, 49]. In particular, the *Mixup* method [53] proposed to generate new training samples using convex combinations of pairs of train-

ing samples and their labels. In order to address the small sample size issue when aligning distributions, we generalize their approach to the semi-supervised learning by using pseudo-labels for unlabeled samples in the interpolation process. Moreover, we also show that by interpolating between labeled and unlabeled data, the empirical distribution of generate data actually gets closer to the unlabeled samples.

3. Problem Statement and Motivations

In semi-supervised learning, we are given a small amount of labeled training samples and a large set of unlabeled training samples. Formally, let us denote by $\mathcal{D}_l = \{(\mathbf{x}_1^l, y_1), \ldots, (\mathbf{x}_n^l, y_n)\}$ as the set of labeled training data, where $\mathbf{x}_i^l$ is the i-th sample, y_i is its corresponding label, and n is the total number of labeled samples. Similarly, the set of unlabeled training data can be represented as $\mathcal{D}_u = \{\mathbf{x}_1^u, \ldots, \mathbf{x}_m^u\}$ where $\mathbf{x}_i^u$ is the i-th unlabeled training sample, and m is the number of unlabeled samples. Usually n is a small number, and we have $m \gg n$. The task of semi-supervised learning is to train a classifier which performs well on the test data drawn from the same distribution with the training data.

3.1. Empirical Distribution Mismatch in SSL

In semi-supervised learning, the labeled training samples $\mathcal{D}_l$ and unlabeled training samples $\mathcal{D}_u$ are assumed to be drawn from an identical distribution. However, due to the limited number of labeled training samples, a considerable difference of empirical distributions can often be observed between the labeled and unlabeled training samples.

More concretely, we take the two-moon data as an example to illustrate the empirical distribution mismatch problem in Figure 1. In particular, the $1,000$ unlabeled samples well describe the underlying distribution (bottom middle), while the labeled samples can hardly represent the two-moon distribution (bottom left). This can be further verified by their distribution by projecting to the x-axis (upper left and upper middle), from which we observe an obvious distribution difference. Actually, when performing multiple rounds of sampling on labeled samples, the empirical distribution of labeled data varies significantly, due to the small sample number.

This phenomenon was also discussed as the sampling bias problem in the literature [23, 24, 30]. In particular, Greton *et al.* [23] pointed out that the difference between two samplings measured by Maximum Mean Discrepancy(MMD) depends on their sampling sizes. In semi-supervised learning where the underlying distribution of labeled and unlabeled data is assumed identical, the MMD of labeled and unlabeled data tends to vanish if and only if both sizes of two sampling are large, which is described as follows,
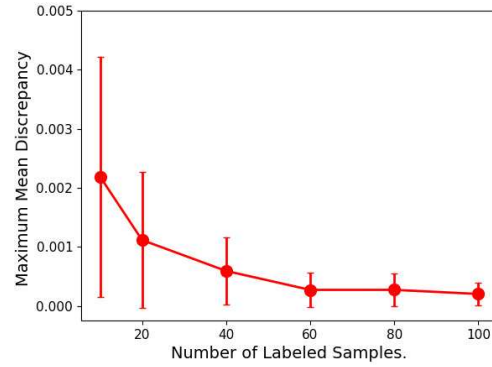

Figure 2. MMD between labeled and unlabeled samples in two-moon example with varying number of labeled samples. Number of unlabeled sample is fixed as $1,000$.

Proposition 1. *Let us denote $\mathcal{F}$ as a class of witness functions $f : \mathbf{x} \to \mathcal{R}$ in the reproduced kernel Hilbert space (RKHS) induced by a kernel function $k(\cdot, \cdot)$, and assume $0 \leq k(\cdot, \cdot) \leq K$, then the MMD distance of $\mathcal{D}_l$ and $\mathcal{D}_u$ can be bounded by $Pr\{MMD[\mathcal{F}, \mathcal{D}_l, \mathcal{D}_u] > 2(\sqrt{(K/n)} + \sqrt{(K/m)} + \epsilon)\} \leq 2\exp\frac{-\epsilon^2 nm}{2K(n+m)}$,*

Proof. The proof can be derived with Theorem 7 in [23] by assuming the two distributions p and q are identical. □

In semi-supervised learning, the number of labeled samples n is usually small, which would lead to a notable empirical distribution difference with the unlabeled samples as stated in above proposition. Specifically, we illustrate the sampling bias problem with the two-moon data in the semi-supervised learning scenario in Figure 2. We plot the MMD between labeled and unlabeled samples with regard to different numbers of labeled samples. As shown in the figure, when the sample size of labeled data is small, both the mean and variance of MMD are large, and the MMD tends to be minor only when n becomes sufficiently large.

This implies that in SSL the small sampling size often causes the empirical approximation of labeled data deviates from the true sample distribution. Consequentially, a model trained from this empirical distribution is unlikely to generalize well on the test data. While various strategies have been exploited for utilizing the unlabeled data in conventional SSL methods [33, 11, 10], the empirical distribution mismatch issue was rarely discussed, which is one of the hidden factors of potentially unstable problem for conventional SSL methods. This was also verified by the recent work [39], which shows that the performance of SSL methods could be degraded when the size of labeled dataset is decreased.

3.2. Healing the Empirical Distribution Mismatch

To overcome the empirical distribution mismatch issue in SSL, in this work, we propose an augmented distribution alignment approach. In addition to training the classifier with supervision from labeled data, we also simultaneously minimize the distribution divergence between labeled and unlabeled data, such that the empirical distributions of labeled and unlabeled samples are well aligned in the latent space (as illustrated in upper right of Figure 1).

Formally, let us denote the loss function as $\ell(f(\mathbf{x}_i^l), y_i)$ where f is the classifier to be learnt. We also define $\Omega(\mathcal{D}_l, \mathcal{D}_u)$ as the distribution divergence of labeled and unlabeled data measured with certain metric. Then, our main idea can be formulated as the following objective,

$$\min_f \sum_{i=1}^{n} \ell(f(\mathbf{x}_i^l), y_i) + \gamma \Omega(\mathcal{D}_l, \mathcal{D}_u), \qquad (1)$$

where γ is a trade-off parameter to balance two terms.

An issue with the above solution is that the small number of labeled samples (*i.e.*, n) potentially makes the optimization of (1) unstable. To address this issue, we further propose a data augmentation strategy. Inspired by the recent mixup approach for supervised learning, we iteratively generate new training samples by interpolating between the labeled samples and unlabeled samples, and feed them for both learning the classifier and reducing the empirical distribution divergence. We refer to our approach as Augmented Distribution Alignment, and detail it in the following section.

4. Augmented Distribution Alignment for SSL

In this section, we introduce our augmented distribution alignment method for SSL, in which we respectively propose two strategies, *adversarial distribution alignment* and *cross-set sample augmentation*, to tackle the empirical distribution mismatch and the small sample issues.

4.1. Adversarial Distribution Alignment

We employ $\mathcal{H}$-Divergence [4, 12] to measure distribution divergence Ω as inspired by recent domain adaptation works.

In particular, let us denote by $g(\cdot)$ a feature extractor (*e.g.*, convolutional layers) which maps sample $\mathbf{x}$ into a latent feature space. Moreover, let $h : g(\mathbf{x}) \to \{0, 1\}$ be a binary discriminator which predicts 0 for labeled samples and 1 for unlabeled samples. The $\mathcal{H}$-Divergence between labeled and unlabeled samples can be written as:

$$d_{\mathcal{H}}(\mathcal{D}_l, \mathcal{D}_u) = 2 \left\{ 1 - \min_{h \in \mathcal{H}} [err(h, g, \mathcal{D}_l) + err(h, g, \mathcal{D}_u)] \right\},$$

where $err(h, g, \mathcal{D}_l) = \frac{1}{n} \sum_{\mathbf{x}^l} [h(g(\mathbf{x}^l)) \neq 0]$ is the prediction error of the discriminator h on labeled samples, and $err(h, g, \mathcal{D}_u)$ is similarly defined for unlabeled samples.

Intuitively, when the empirical distribution mismatch is large, the discriminator could easily distinguish the labeled and unlabeled samples, thus its prediction errors would be small, and the $\mathcal{H}$-divergence is higher, and vice versa. Therefore, to reduce the empirical distribution mismatch of labeled and unlabeled samples, we then minimize the distribution distance $d_{\mathcal{H}}(\mathcal{D}_l, \mathcal{D}_u)$ to enforce the feature extractor g to generate a latent space in which two sets of features are well aligned. This is therefore achieved by solving the following problem:

$$\min_g d_{\mathcal{H}}(\mathcal{D}_l, \mathcal{D}_u) = \max_g \min_{h \in \mathcal{H}} [err(h, g, \mathcal{D}_l) + err(h, g, \mathcal{D}_u)]$$

The above max-min problem can be optimized with the adversarial training methods. In [17], Ganin and Lempitsky showed that it can be implemetned as a simple gradient reverse layer (GRL) which automatically reverse the gradient after discriminator, thus one can directly minimize the classification loss of the discriminator h with the standard propagation optimization library.

4.2. Cross-set Sample Augmentation

As discussed in Section 3, in SSL, the limited sampling size of labeled data often causes unstable in optimization and leads to performance degradation. In order to reinforce the alignment, as inspired by [53], we propose to generate new training samples by interpolating between labeled and unlabeled samples. In particular, for each $\mathbf{x}^u$, we assign it a pseudo-label $\hat{y}^u$, which is generated by using the prediction from the model trained in previous iteration in this work. Then, given a labeled sample $\mathbf{x}^l$ and an unlabeled sample $\mathbf{x}^u$, the interpolated sample can be represented as,

$$\tilde{\mathbf{x}} = \lambda \mathbf{x}^l + (1 - \lambda)\mathbf{x}^u, \qquad (2)$$
$$\tilde{y} = \lambda y^l + (1 - \lambda)\hat{y}^u, \qquad (3)$$
$$\tilde{z} = \lambda \cdot 0 + (1 - \lambda) \cdot 1, \qquad (4)$$

where λ is a random variable that is generated from an prior β distribution, *i.e.* $\lambda \sim \beta(\alpha, \alpha)$ with α being a hyperparameter to control the shape of the β distribution, $\tilde{\mathbf{x}}$ is the interpolated sample, $\tilde{y}$ is its class label, and $\tilde{z}$ is its label for the distribution discriminator.

The benefits of such cross-set sample augmentation are two-fold. First, the interpolated samples greatly enlarged the training data set, making the learning process more stable, especially for deep neural networks models. It was also shown in [53] that such data augmentation helps to improve model robustness.

Second, each pseudo-sample is generated by interpolating between a labeled sample and an unlabeled sample, thus
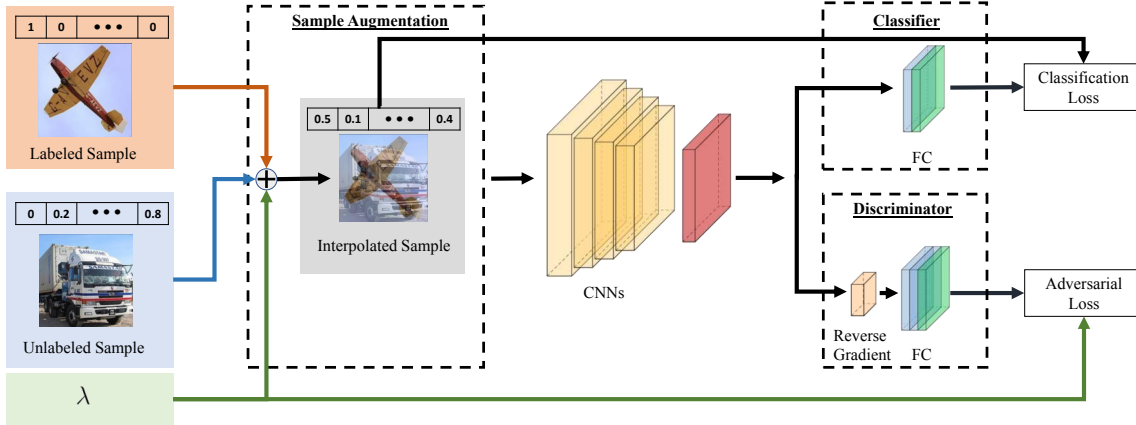
Figure 3. The network architecture of our proposed ADA-Net, in which we append an additional discriminator classifier branch with a gadient reverse layer to the vanilla CNN (shown in the bottom right part). In training time, the cross-set sample interpolation is performed between labeled and unlabeled samples, and we feed the interpolated samples into the network. Pesudo-labels of unlabeled samples are obtained using the classifier trained in last iteration (see explanation in Section 4.3) for details.

the distribution of pseudo-samples is expected to be closer to the real distribution than that of the original labeled training samples. We prove this using the euclidean generalized energy distance [46] in below.

Let us denote P_l and P_u as the empirical distributions of labeled and unlabeled data, their euclidean generalized energy distance [46] can be written as,

$$J^2(P_l, P_u) = \mathbb{E}[\|\mathbf{x}^l - \mathbf{x}^u\|^2] - \mathbb{E}[\|\mathbf{x}^l - \mathbf{x}^{l'}\|^2] - \mathbb{E}[\|\mathbf{x}^u - \mathbf{x}^{u'}\|^2].$$

where $\| \cdot \|$ is the euclidean distance, $\mathbf{x}^l$ and $\mathbf{x}^{l'}$ (*resp.*, $\mathbf{x}^u$ and $\mathbf{x}^{u'}$) are two samples independent sampled from P_l (*resp.*, P_u). Then, we show that cross-set sample augmentation helps to bridge the gap between two distributions by the following proposition,

Proposition 2. *Let $\tilde{P}$ be the empirical distribution of the pseudo sample $\tilde{\mathbf{x}}$ generated using (2), then we have $J^2(\tilde{P}, P_u) = \frac{1}{4} J^2(P_l, P_u)$. In other words, the euclidean generalized energy distance between the empirical distribution of the pseudo and unlabeled samples is smaller or equal than that of labeled and unlabeled samples.*

Proof. Using Proposition 2 from [46], we rewrite the energy distance $J^2(P_l, P_u)$ as follows,

$$J^2(P_l, P_u) = 2\|\mathbb{E}[\mathbf{x}^l] - \mathbb{E}[\mathbf{x}^u]\|^2$$

In addition, we have

$$\mathbb{E}[\lambda \mathbf{x}^l + (1 - \lambda)\mathbf{x}^u] = \frac{1}{2}\mathbb{E}[\mathbf{x}^l] + \frac{1}{2}\mathbb{E}[\mathbf{x}^u],$$

because the expectation of $\lambda \sim \beta(\alpha, \alpha)$ is 0.5, and the same

applies to $1 - \lambda$. Therefore,

$$\begin{aligned} J^2(\tilde{P}, P_u) &= 2\|\frac{1}{2}\mathbb{E}[\mathbf{x}^l] + \frac{1}{2}\mathbb{E}[\mathbf{x}^u] - \mathbb{E}[\mathbf{x}^u]\|^2 \\ &= 2\|\frac{1}{2}\mathbb{E}[\mathbf{x}^l] - \frac{1}{2}\mathbb{E}[\mathbf{x}^u]\|^2 \\ &= \frac{1}{4}J^2(P_l, P_u) \end{aligned}$$

Here we complete the proof. □

This implies that the new generated pseudo-samples can be deemed as being sampled from the intermediate distributions between the empirical distributions of labeled and unlabeled samples. As shown in previous domain adaptation works [21, 18], such intermediate distributions are beneficial to alleviate the gap between two distributions, and learn more robust models.

4.3. Summary

We unify the adversarial distribution alignment and cross-set sample augmentation strategies into one framework, finally leading to our augmented distribution alignment approach.

In Figure 3, we demonstrate an example of incorporating our augmented distribution alignment approach into a vanilla convolutional neural networks, which is referred to as *ADA-Net*. Specifically, in addition to the classification branch, we add several fully connected layers as the discriminator to distinguish labeled and unlabeled samples (*i.e.*, h discussed in Section 4.1). A gradient reverse layer is added before the discriminator, which will automatically reverse the sign of gradient from the discriminator during

Algorithm 1: A training step for ADA-Net.

Input : A batch of labeled samples $\{(\mathbf{x}^l, y^l), \ldots\}$, a batch of unlabeled samples $\{\mathbf{x}^u, \ldots\}$, classifier f and discriminator h.

1. Run one forward step to get pseudo-labels for unlabeled samples, *i.e.*, $\hat{y}^u \leftarrow f(\mathbf{x}^u)$

2. Sample λ of batch size from $\beta(\alpha, \alpha)$, and generate a batch of samples $\{(\tilde{\mathbf{x}}, \tilde{y}, \tilde{z}), \ldots\}$ using (2),(3),(4).

3. Perform a forward pass by feeding $\{(\tilde{\mathbf{x}}, \tilde{y}, \tilde{z}), \ldots\}$.

4. Perform a backward pass by minimizing (5).

Output: classifier f and discriminator h

---

back-propagation. Then, for each mini-batch, we use the cross-set sample augmentation strategy in (2),(3),(4) to generate interpolated samples and labels, and use them as training data to train our ADA-Net. The objective for training the network can be obtained by replacing the training samples and $\Omega(\cdot, \cdot)$ term in (1), *i.e.*,

$$\min_{f,g,h} \sum_{\tilde{\mathbf{x}}} \lambda \ell(f(g(\tilde{\mathbf{x}})), \tilde{y}) + \gamma \ell(h(g(\tilde{\mathbf{x}})), \tilde{z}), \qquad (5)$$

where g, f, h are respectively the feature extractor, classifier, and discriminator, and $\ell(\cdot, \cdot)$ is the loss function for which we use the cross-entropy in this work. λ is the weight for classification loss, which corresponds to the λ for generating the interpolated sample $\tilde{\mathbf{x}}$ (see (2)). The reason for applying this weight is as follows. The higher λ is, the higher proportion of $\tilde{\mathbf{x}}$ coming from labeled set is, and we are more confident on its label $\tilde{y}$, and vice versa.

We depict the training pipeline Algorithm 1. Aside from the simple sample interpolation, the network can be optimized with the standard propagation approaches. Therefore, our augmented distribution alignment can be easily incorporated into existing neural networks by appending a discriminator with the GRL layer, and adding the proposed cross-set sample augmentation during mini-batch data preparation.

5. Experiments

In this section, we evaluate our proposed ADA-Net for semi-supervised learning on benchmark datasets including SVHN, and CIFAR10.

5.1. Experimental Setup

SVHN: The Street View House Numbers (SVHN) dataset [38] is a dataset consists of real-world digit photos. It includes ten classes and 73,257 training images of 32×32

size. Following [37], out of the full training set, 1000 images are used with labels for supervised learning. The rest training photos are provided without labels. Random translation is the only augmentation used for this dataset.

CIFAR10: The CIFAR10 dataset [31] contains 10 classes, and consists of 50,000 training images as well as 10,000 test images. All images are of the size 32×32. 4,000 samples from the training images are used as labeled set for our experiments, the rest are used as unlabeled samples.

We use the PreAct-ResNet-18 [26] as the backbone network, and implement our ADA-Net in Tensorflow based on the open source TensorPack library [50]. For the class classifier, a single fully connected layer is used to map the features to logits. For the domain classifier, two dense layers, each with 1,024 units, followed by another dense layer are used to produce two channels of soft domain labels.

The batch size is set as 128. The learning rate starts from 0.1, and is divided by 10 when 50%, and 75% epochs are reached. The network is trained for 100 epochs in total for SVHN, and 300 epochs for CIFAR10, where one epoch is defined as one iteration over all unlabeled data. We use a momentum optimizer with 0.9 as the momentum. The following hyperparameters are used for our reported results: weight-decay= 0.0001, interpolation $\alpha = 0.1$ for SVHN and $\alpha = 1.0$ for CIFAR10. The experiments on SVHN and CIFAR10 share the exact same network and protocol. We followed [37] to use a separate validation set of 1,000 images to select α for all methods.

5.2. Experimental Results

We summarize the classification error rates on the SVHN and CIFAR10 dataset in Table 1. We include the baseline CNN model that is trained with labeled data only as a reference. To validate the effectiveness of the two modules in our ADA-Net, we also report two variants of our proposed approach. In the first variant, we do not use cross-set sample augmentation and apply the distribution alignment using original labeled and unlabeled samples. In the second variant, we remove the discriminator and perform only cross-set sample augmentation for learning the classifier.

As shown in Table 1, our ADA-Net significantly improves the classification performance on both datasets. We also observe that both the distribution alignment and cross-set sample augmentation are important for improving the performance. The distribution alignment module brings 1.30% and 3.04% improvement on CIFAR10 and SVHN, and the cross-set sample augmentation module gives 6.18% and 3.06% improvement, respectively. By integrating both modules, the classification error rates can be reduced by our ADA-Net from 19.97% and 13.80% to 8.87% and 5.90% on the CIFAR10 and SVHN datasets, respectively. The experimental results clearly validate our motivations, and also
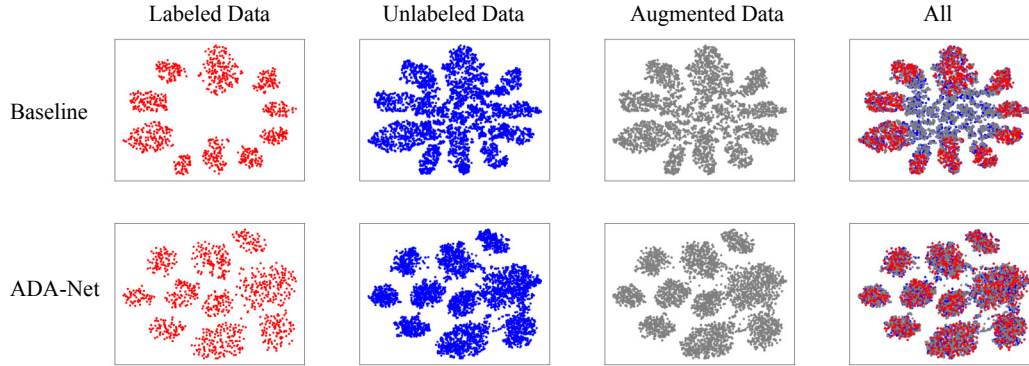
Figure 4. Visualization of SVHN features obtained by baseline CNN and our ADA-Net using t-SNE. We generated the t-SNE using labeled, unlabeled, and interpolated samples together, and show them separately for a better comparison. For baseline CNN, empirical distribution mismatch between labeled and unlabeled samples can be observed, and the augmented samples bridge the gap to some extent. For our ADA-Net, with the augmented distribution alignment, empirical distribution mismatch are well reduced.

demonstrate the effectiveness of our proposed approach.

5.3. Experimental Analysis

Feature visualization: To better understand how our ADA-Net works, we use the base CNN block as a feature extractor, and visualize with the t-SNE approach for the labeled samples, unlabeled samples, and the generated pseudo-samples on the SVHN dataset in Figure 4. The features extracted using the baseline CNN trained with only labeled data are also visualized for comparison. As shown in Figure 4, a considerable distribution difference between labeled and unlabeled samples can be observed for the baseline CNN model, and the generated pseudo-samples distribute in between those two sets. Nevertheless, with our ADA-Net, the distributions of three types of samples are similar since we explicitly align the distributions of labeled and unlabeled samples in the training procedure.

Table 1. Classfication error rates of our proposed ADA-Net and its variants on the CIFAR10 and SVHN datasets. "dist" denotes the distribution alignment module, and "aug" denotes the cross-set sample augmentation module. PreAct-ResNet-18 [26] is used as the backbone network.

	dist	aug	CIFAR10	SVHN
Baseline			19.97%	13.80%
Ours	✓		18.67%	10.76%
		✓	13.79%	10.74%
	✓	✓	**8.87%**	**5.90**%

Feature distribution: To further show the effectiveness of our ADA-Net in reducing the distribution mismatch, we take the first three activations of the baseline CNN model and our ADA-Net as examples, and plot the distribution of labeled and unlabeled samples on each dimension individually. The distribution is obtained by performing kernel density estimation [40, 43] on each type of samples and each

dimension individually. As shown in Figure 6, we again observe a considerable mismatch between the estimated empirical distribution of labeled and unlabeled samples for the baseline CNN model. And also, such distribution mismatch is then well reduced in our ADA-Net model. We have similar observations for other feature activations.
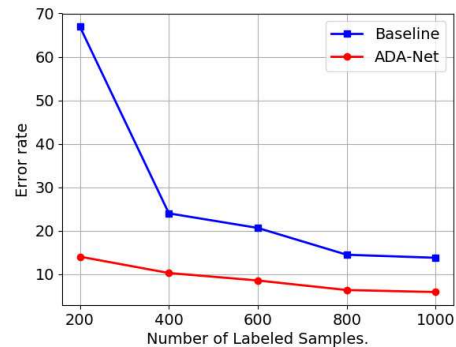

Figure 5. Classification Error rates on SVHN of our ADA-Net and baseline CNN when varying the number of labeled samples.

Varying number of labeled samples: As discussed in Section 3.1, the distribution mismatch in SSL is correlated with the number of labeled samples. It often becomes more serious when the number of labeled samples is less. To validate the effectiveness of ADA-Net with different sampling size, we conduct experiments on the SVHN dataset by varying the number of labeled samples. In particular, we train models using 200, 400, 600, 800 and $1,000$ labeled samples, and all other experimental settings remain the same. The error rates of ADA-Net and the baseline CNN are plotted in Figure 5. We observe that the error rate of baseline CNN model increases dramatically when reducing the number of labeled samples, which indicates that the sampling bias makes the learning problem more challenging. Nevertheless, our ADA-Net consistently improves the classifica-
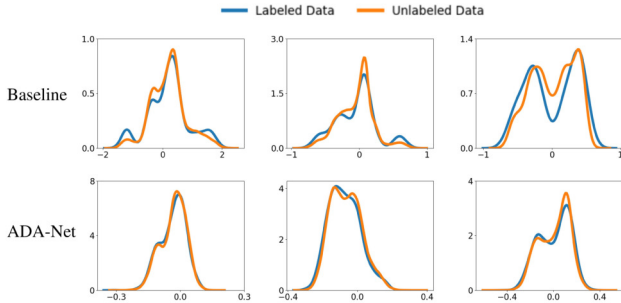
Figure 6. Kernel density estimation of labeled and unlabeled samples of the SVHN Dataset based on the first three feature activations of the baseline CNN and our ADA-Net. Considerable distribution mismatch between labeled and unlabeled data can be observed for the baseline CNN model (top row), while two distributions are generally aligned well with our ADA-Net (bottom row).

tion performance by alleviating such sampling bias with the augmented distribution alignment, the relative improvement is more obvious when the labeled samples are rare.

5.4. Comparison with State-of-the-arts

Table 2. Classification error rates of different methods on CIFAR10 and SVNH. Conv-Large [47] is used as the backbone network. Results of baseline methods are taken from the papers.

Method	CIFAR10	SVHN
Π Model [32]	12.36%	4.82%
Temporal ensembling [32]	12.16%	4.42%
Mean Teacher[47]	12.31%	3.95%
VAT [37]	11.36%	5.42%
VAT+Ent [37]	10.55%	3.86%
SaaS [11]	13.22%	4.77%
MA-DNN [10]	11.91%	4.21%
VAT+Ent+SNTG [35]	9.89%	3.83%
Mean Teacher+fastSWA* [2]	9.05%	-
ADA-Net (Ours)	10.30%	4.62%
ADA-Net+ (Ours)	10.09%	**3.54%**
ADA-Net*(Ours)	**8.72%**	-

* Larger translation range (4 instead of 2), and without ZCA whitening.

Table 3. Classification error rates of different methods on ImageNet dataset. ResNet-18 is used as the backbone network.

Method	Top-1	Top-5
100% Supervised	30.43%	10.76%
10% Supervised	52.23%	27.54%
Mean Teacher [47]	49.07%	23.59%
Dual-View Deep Co-Training [41]	46.50%	22.73%
ADA-Net (Ours)	**44.91**%	**21.18**%

We further compare our ADA-Net with recently proposed state-of-the-art SSL learning approaches [32, 47, 37, 37, 11, 10, 35, 2]. As discussed in [39], minor modification in the network structure and data processing method often

lead to different results. To ensure a fair comparison, we take the VAT method [37] as a reference, and strictly follow their experimental setup. In particular, we re-implement our ADA-Net based on the released codes from [37]. The same Conv-Large architecture and hyper-parameters are used.

We report the results of different methods on the CIFAR10 and SVHN datasets in Table 2. Our ADA-Net achieves competitive results with those state-of-the-art SSL methods. Despite the simplicity of our augmented distribution alignment, the results clearly validate the importance on dealing with the empirical distribution mismatch in the semi-supervised learning, and also demonstrates the effectiveness of our ADA-Net. Furthermore, by adopting a simpler augmentation setup used by [2], our vanilla ADA-Net approach pushes the envelope of SSL on CIFAR10, and achieves new state-of-the-art error rates of 8.72%.

More importantly, as we solve the SSL problem from a new perspective that was not revealed by previous works, our augmented distribution alignment strategy is generally complementary to other methods. Therefore, the performance of existing SSL methods can be boosted by incorporating the distribution alignment and cross-set sample augmented modules proposed in this work. As shown in Table 2, combining our ADA-Net with the VAT+Ent method (denoted as "ADA-Net+"), we achieve new state-of-the-art error rates of 3.54% on SVHN.

We additionally report our result on 1000-class ImageNet in Table 3, with 10% labels. We compare our results with previous state-of-the-art methods Mean Teacher [47] and Deep Co-Training [41]. The result of Deep Co-Training is quoted from their paper, and the performance of Mean Teacher is from running their official implementation by [41]. Following [41], we train ResNet-18 [25] for 600 epochs with a batch size of 256, and we set $\alpha = 1.0$. ADA-Net performs better than both methods and outperforms Dual-View Deep Co-Training by 1.59% on Top-1 error rate and 1.55% on Top-5 error rate.

6. Conclusions

In this work, we have proposed a new semi-supervised learning method called augmented distribution alignment. In particular, we tackle the semi-supervised learning problem from a new perspective that labeled and unlabeled data often exhibits a considerable difference in terms of the empirical distribution. We therefore employed an adversarial training strategy to align the distributions of labeled and unlabeled samples when training the neural networks. A cross-set sample augmentation was further proposed to deal with the limited sampling size and bridge the distribution gap. Those two strategies can be readily unified into the existing deep neural networks, leading to our ADA-Net. Experiments on the benchmark CIFAR10 and SVHN datasets have validated the effectiveness of our approach.

References

[1] Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(Nov):1817–1853, 2005. 2

[2] Ben Athiwaratkun, Marc Finzi, Pavel Izmailov, and Andrew Gordon Wilson. There are many consistent explanations of unlabeled data: Why you should average. In *International Conference on Learning Representations*, 2019. 8

[3] Mikhail Belkin and Partha Niyogi. Semi-supervised learning on riemannian manifolds. *Machine learning*, 56(1-3):209–239, 2004. 1

[4] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010. 4

[5] Avrim Blum and Shuchi Chawla. Learning from labeled and unlabeled data using graph mincuts. 2001. 1, 2

[6] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM, 1998. 1, 2

[7] Olivier Chapelle, Jason Weston, and Bernhard Schölkopf. Cluster kernels for semi-supervised learning. In *Advances in neural information processing systems*, pages 601–608, 2003. 1

[8] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1841–1850, 2019. 2

[9] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018. 2

[10] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Semi-supervised deep learning with memory. In *The European Conference on Computer Vision (ECCV)*, 2018. 1, 2, 3, 8

[11] Safa Cicek, Alhussein Fawzi, and Stefano Soatto. SaaS: Speed as a supervisor for semi-supervised learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 149–163, 2018. 1, 2, 3, 8

[12] Corinna Cortes and Mehryar Mohri. Domain adaptation in regression. In *International Conference on Algorithmic Learning Theory*, pages 308–323. Springer, 2011. 4

[13] Debasmit Das and CS George Lee. Unsupervised domain adaptation using regularized hyper-graph matching. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3758–3762. IEEE, 2018. 2

[14] Miroslav Dudik, Steven J Phillips, and Robert E Schapire. Correcting sample selection bias in maximum entropy density estimation. In *Advances in neural information processing systems*, pages 323–330, 2006. 2

[15] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE international conference on computer vision*, pages 2960–2967, 2013. 2

[16] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189, 2015. 2

[17] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. 4

[18] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2066–2073. IEEE, 2012. 5

[19] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. Dlow: Domain flow for adaptation and generalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2477–2486, 2019. 2

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[21] Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *2011 international conference on computer vision*, pages 999–1006. IEEE, 2011. 5

[22] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, pages 529–536, 2005. 1, 2

[23] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012. 3

[24] Arthur Gretton, Kenji Fukumizu, Zaid Harchaoui, and Bharath K Sriperumbudur. A fast, consistent kernel two-sample test. In *Advances in neural information processing systems*, pages 673–681, 2009. 3

[25] 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. 8

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016. 6, 7

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

[28] Hiroshi Inoue. Data augmentation by pairing samples for images classification. *arXiv preprint arXiv:1801.02929*, 2018. 2

[29] Thorsten Joachims. Transductive learning via spectral graph partitioning. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, pages 290–297, 2003. 2

[30] Sudeep Kamath, Alon Orlitsky, Dheeraj Pichapati, and Ananda Theertha Suresh. On learning distributions from their samples. In *Conference on Learning Theory*, pages 1066–1100, 2015. 3

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

[32] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 1, 2, 8

[33] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, page 2, 2013. 1, 3

[34] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pages 97–105, 2015. 2

[35] Yucen Luo, Jun Zhu, Mengxi Li, Yong Ren, and Bo Zhang. Smooth neighbors on teacher graphs for semi-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8896–8905, 2018. 8

[36] Tom M Mitchell. The role of unlabeled data in supervised learning. In *Language, Knowledge, and Representation*, pages 103–111. Springer, 2004. 2

[37] Takeru Miyato, Shin-ichi Maeda, Shin Ishii, and Masanori Koyama. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 1, 2, 6, 8

[38] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 5, 2011. 6

[39] Augustus Odena, Avital Oliver, Colin Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of semi-supervised learning algorithms. 2018. 1, 2, 3, 8

[40] Emanuel Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962. 7

[41] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. Deep co-training for semi-supervised image recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 135–152, 2018. 8

[42] Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. Semi-supervised self-training of object detection models. 2005. 2

[43] Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, pages 832–837, 1956. 7

[44] Saharon Rosset, Ji Zhu, Hui Zou, and Trevor J Hastie. A method for inferring label sampling mechanisms in semi-supervised learning. In *Advances in neural information processing systems*, pages 1161–1168, 2005. 2

[45] Baochen Sun and Kate Saenko. Subspace distribution alignment for unsupervised domain adaptation. In *BMVC*, pages 24–1, 2015. 2

[46] Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013. 5

[47] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204, 2017. 1, 2, 8

[48] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017. 2

[49] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Aaron Courville, Ioannis Mitliagkas, and Yoshua Bengio. Manifold mixup: Learning better representations by interpolating hidden states. 2018. 2

[50] Yuxin Wu et al. Tensorpack, 2016. 6

[51] Yuguang Yan, Wen Li, Michael KP Ng, Mingkui Tan, Hanrui Wu, Huaqing Min, and Qingyao Wu. Learning discriminative correlation subspace for heterogeneous domain adaptation. In *IJCAI*, pages 3252–3258, 2017. 2

[52] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, 1995. 2

[53] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 2, 4

[54] Weichen Zhang, Wanli Ouyang, Wen Li, and Dong Xu. Collaborative and adversarial network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3801–3809, 2018. 2

[55] Xiaojin Jerry Zhu. Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2005. 1, 2