Towards Robust Cross-domain Image Understanding with Unsupervised Noise Removal

Lei Zhu
National University of Singapore
zhu-lei@comp.nus.edu.sg

Zhaojing Luo*
National University of Singapore
zhaojing@comp.nus.edu.sg

Wei Wang
National University of Singapore
wangwei@comp.nus.edu.sg

Meihui Zhang
Beijing Institute of Technology
meihui_zhang@bit.edu.cn

Gang Chen
Zhejing University
cg@zju.edu.cn

Kaiping Zheng
National University of Singapore
kaiping@comp.nus.edu.sg

ABSTRACT

Deep learning has made a tremendous impact on various applications in multimedia, such as media interpretation and multimodal retrieval. However, deep learning models usually require a large amount of labeled data to achieve satisfactory performance. In multimedia analysis, domain adaptation studies the problem of cross-domain knowledge transfer from a label rich source domain to a label scarce target domain, thus potentially alleviates the annotation requirement for deep learning models. However, we find that contemporary domain adaptation methods for cross-domain image understanding perform poorly when source domain is noisy. Weakly Supervised Domain Adaptation (WSDA) studies the domain adaptation problem under the scenario where source data can be noisy. Prior methods on WSDA remove noisy source data and align the marginal distribution across domains without considering the fine-grained semantic structure in the embedding space, which have the problem of class misalignment, e.g., features of cats in the target domain might be mapped near features of dogs in the source domain. In this paper, we propose a novel method, termed Noise Tolerant Domain Adaptation (NTDA), for WSDA. Specifically, we adopt the cluster assumption and learn cluster discriminatively with class prototypes (centroids) in the embedding space. We propose to leverage the location information of the data points in the embedding space and model the location information with a Gaussian mixture model to identify noisy source data. We then design a network which incorporates the Gaussian mixture noise model as a sub-module for unsupervised noise removal and propose a novel cluster-level adversarial adaptation method based on the Generative Adversarial Network (GAN) framework which aligns unlabeled target data with the less noisy class prototypes for mapping the semantic structure across domains. Finally, we devise a simple and effective algorithm to train the network from end to end. We conduct extensive experiments to evaluate the effectiveness of our method on both general images and medical images from COVID-19 and e-commerce datasets. The results show that our method significantly outperforms state-of-the-art WSDA methods.

CCS CONCEPTS

• **Computing methodologies** → **Transfer learning**; *Image representations*; • **Computer systems organization** → **Neural networks**.

KEYWORDS

Representation Learning; Weakly Supervised Domain Adaptation; Adversarial Learning

ACM Reference Format:
Lei Zhu, Zhaojing Luo, Wei Wang, Meihui Zhang, Gang Chen, and Kaiping Zheng. 2021. Towards Robust Cross-domain Image Understanding with Unsupervised Noise Removal. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3474085.3475175

1 INTRODUCTION

There is great interest in using deep learning for various multimedia applications, such as media interpretation [9, 35], multimodal retrieval [15, 37–39]. Much of its success is attributed to the availability of large-scale labeled training data [7]. However, in practice, large-scale labeled data are hardly available, as manual annotating sufficient label information for various multimedia applications is both expensive and time-consuming. Thus, it is desirable to reuse labeled data from a related domain for cross-domain image understanding. This process is called Domain Adaptation (DA), which transfers knowledge from a label rich source domain to a label scarce target domain [24]. Intuitively, the data quality in the source domain affects the domain adaptation performance. However, in practice, high-quality source data related to a target task of interest is hardly available. In contrast, the Internet and social media contain large-scale labeled multimedia data which can be downloaded with keyword search [14, 40], but unfortunately, these data contain noise, either in features, labels or both. Similarly in medical image analysis, annotating medical data requires medical expertise and due to subjectivity of domain experts and diagnostic difficulties, noisy labels are often inevitable. Thus, it is meaningful to study robust domain adaptation under the scenario when source data is noisy in order for better cross-domain image understanding. This problem has been referred to as Weakly Supervised Domain Adaptation (WSDA) [29].

*Corresponding author.

MM '21, October 20–24, 2021, Virtual Event, China.

ACM ISBN 978-1-4503-8651-7/21/10.
https://doi.org/10.1145/3474085.3475175

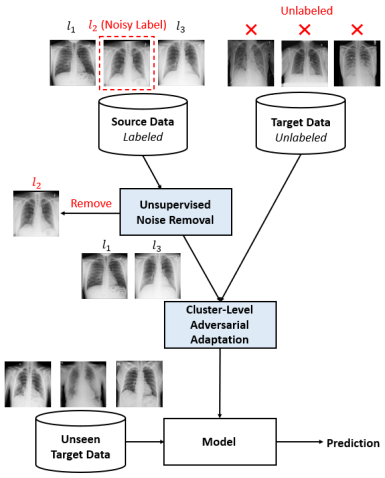

Figure 1: Workflow of Noise Tolerant Domain Adaptation for medical image analysis. Source and Target data are chest X-ray images acquired with different scanners for lung disease diagnosis.

Although WSDA enables many practical use cases of domain adaptation in real life and can substantially reduce the annotation costs, it is still not well studied in the literature. There are two entangled challenges in WSDA, namely source data noise and distribution shift across domains. Directly applying existing domain adaptation methods for WSDA will not work, as source data noise severely deteriorates the adaptation performance [16, 29]. Liu et al. [16] recently propose a Butterfly framework to address these issues. However, their work only considers the scenario where source domain contains label noise data and cannot handle feature noise data. Another limitation of the method is its large model size, and as a result, it incurs a large amount of computational resources to train the model.

Shu et al. [29] recently propose transferable curriculum for WSDA. The transferable curriculum can select transferable and clean source data for adversarial domain adaptation, which makes their method robust to source data noise. However, one problem with their method is that the adversarial learning between the feature extractor and domain discriminator only aligns the marginal distribution across domains and ignores the fine-grained class structure in the embedding space. Noisy source data distorts the original source data distribution, thus directly aligning the marginal distribution across domains would potentially transfer the noise information from source domain to target domain and possibly cause class misalignment due to label noise, where even with perfect alignment of the marginal distribution, the class structure across domains may not be well aligned, e.g., features of cats in target domain might be mapped near features of dogs in source domain, which leads to poor target performance [31, 36].

Recently, several Unsupervised Domain Adaptation (UDA) methods [8, 28] adopt the cluster assumption [4] to alleviate the class misalignment problem, where they assume data distributes in the embedding space with separated data clusters and data samples in the same cluster share the same class label. In [28], Shu et al. propose Virtual Adversarial Domain Adaptation (VADA) which combines virtual adversarial training [22] and conditional entropy loss to push the decision boundaries away from class clusters in the embedding space. In [8], Deng et al. propose Cluster Alignment with a Teacher (CAT), which forces features of both source domain and target domain to form discriminative class-conditional clusters and aligns the corresponding clusters across domains. Although both works alleviate the class misalignment problem and demonstrate significant improvement in performance over prior domain adaptation methods that only align marginal distribution across domains, however, like most existing domain adaptation methods, they perform poorly when source domain contains noise [16, 29].

In light of the issues with existing WSDA methods and inspired from recent clustering based domain adaptation methods, in this paper, we propose a novel method for WSDA with unsupervised noise removal to address these issues. For its noise tolerance property, we call our method, Noise Tolerant Domain Adaptation, or NTDA in short. Specifically, we learn the clusters discriminatively with class prototypes (centroids) in the embedding space. We propose to estimate the probability of a source data point being noisy with a Gaussian mixture noise model based on its distance to the class prototype and filter source data points with high probabilities of being noisy (See Sec. 3.2 for more details). We incorporate the Gaussian mixture noise model as a sub-module within a deep network and propose a cluster-level adversarial adaptation method based on the GAN framework which aligns unlabeled target data with the less noisy class prototypes for mapping the semantic structure across domains. Finally, we devise a simple and effective algorithm to train the network from end to end. Fig. 1 presents the workflow of our proposed NTDA method.

To summarize, we make the following contributions in this paper:

- We identify several issues with existing WSDA methods and propose a simple and effective method, NTDA, to address these issues.
- We propose a novel Unsupervised Noise Removal (UNR) method based on the location information of data points, which can be applied to other clustering based domain adaptation methods to make them robust to source data noise.
- We propose a Cluster-Level Adversarial Adaptation (CAA) method, which adversarially aligns target data points with the less noisy class prototypes in the embedding space and alleviates class misalignment problem.
- We conduct extensive experiments to evaluate the effectiveness of our method on both general images and medical images. The results show that NTDA significantly improves state-of-the-art results for WSDA.

NTDA has been developed as part of the library for MLCask [21] for supporting healthcare analytics. MLCask is a model-data provenance and pipeline management system that rides on Apache SINGA [23] for supporting end-to-end analytics. The remainder of the paper is organized as follows. Section 2 provides a brief background on domain adaption, and related works. In Section 3, we present our methodology to tackle WSDA problems and propose NTDA. We conduct extensive experimental study and present the results in Section 4. We conclude in Section 5.

2 RELATED WORKS

Domain Adaptation [24] aims to build models that generalize across domains. Existing domain adaptation methods can be roughly divided into two major categories: discrepancy-based and adversarial-based methods. Discrepancy-based methods align feature distributions across domains by minimizing certain distribution discrepancy, such as Maximum Mean Discrepancy (MMD) [17, 33], correlation distance [30] or Central Moment Discrepancy (CMD) [44]. Adversarial-based methods draws inspiration from the two-player game of Generative Adversarial Networks (GAN) [11]. DANN [10] trains domain invariant features via adding a domain classifier in the deep feature learning pipeline via gradient reversal. ADDA [32] adopts asymmetric feature extractors for adversarial training. CDAN [18] conditions the adversarial domain adaptation models on discriminative information conveyed in the classifier prediction. Our method is especially related to the more recent clustering based domain adaptation methods [8, 28, 46] with cluster assumption.

Learning From Noisy Data is an active research area in machine learning. Recently, Zhang et al. [45] empirically demonstrate that noisy data will be memorized by deep networks which destroys their generalization capability. Arpit et al. [1] find that when training with noisy data, deep networks learn simple patterns first before memorizing noisy data. Based on this memorization effect of deep networks, Han et al. [12] propose a training paradigm termed "co-teaching" where they train two networks simultaneously, and utilize the small loss data from one network to teach the other one, which is called the small-loss trick. Yu et al. [42] further propose the "Update by Disagreement" strategy with "co-teaching" to prevent the two networks converge to consensus.

Weakly Supervised Domain Adaptation (WSDA) studies the domain adaptation problem under the scenario where source data can be noisy. Although WSDA studies a more practical problem than ordinary domain adaptation, it is still under-explored in the literature. Recently, Liu et al. [16] propose a Butterfly framework which consists of four networks for WSDA. However, their method demands a large amount of computational resources for training. Shu et al. [29] propose transferable curriculum to select transferable and clean source data for WSDA, but they ignore the class structure in the embedding space for distribution alignment. Yu et al. [43] propose a theoretical framework for label-noise robust domain adaptation with a denoising Conditional Invariant Component. However, their method cannot handle cases when there are feature noise in the data. Zhang et al. [48] propose Collaborative Unsupervised Domain Adaptation for general and medical image analysis, where they optimize two networks collaboratively to learn from noisy source data and perform weighted instance-level domain adaptation with unlabeled target data. However, their method aligns the marginal distribution across domains to reduce the distribution shift which suffers from class misalignment under the label noise scenario.

3 METHODOLOGY

In Unsupervised Domain Adaptation (UDA), we are given N^s labeled data $\mathbb{D}^s = \{(\boldsymbol{x}_i^s, y_i^s)\}_{i=1}^{N^s}$ in source domain and N^t unlabeled data $\mathbb{D}^t = \{\boldsymbol{x}_i^t\}_{i=1}^{N^t}$ in target domain. The source and target data share the same set of labels and are sampled from probability distributions P^s and P^t respectively with $P^s \neq P^t$. In Weakly Supervised Domain Adaptation (WSDA), we relax the assumption that $\mathbb{D}^s$ is clean to that $\mathbb{D}^s$ may be corrupted with noise in labels, features or both. The goal of WSDA is to effectively transfer knowledge from noisy source domain to unlabeled target domain.

3.1 Preliminary: Discriminative Clustering with Class Information

In this paper, we adopt the cluster assumption [4], where we assume data distribution in the embedding space contains separated data clusters and data samples in the same cluster share the same class label. With labeled source data $\mathbb{D}^s$, we propose to learn the clusters discriminatively with class information. We employ the prototype learning framework from [41] where we assign a prototype for each class in the embedding space. Let $\boldsymbol{f}_i^s = F(\boldsymbol{x}_i^s; \Phi)$ be the feature for source data $\boldsymbol{x}_i^s$ with label y_i^s, $\boldsymbol{p}_j$ be the class prototype for the jth class and $\{\boldsymbol{p}_j\}_{j=1}^M$ be the set of prototypes, where $F : \mathcal{X} \rightarrow \mathbb{R}^d$ is the feature extractor, d is the embedded feature dimension, Φ is the set of parameters for F and M is the total number of classes. We measure the probability of a data point belonging to a specific class with a softmax over distance to prototypes as shown in Eqn. 1, where $d(\boldsymbol{f}, \boldsymbol{p}_j) = ||\boldsymbol{f} - \boldsymbol{p}_j||_2^2$ is the squared euclidean distance between two data points in the embedding space and T is the hyper-parameter for scaling the exponent value. To train the network, we employ cross entropy loss on the prediction probability as shown in Eqn. 2.

$$P(y|\boldsymbol{f}) = \frac{e^{-\frac{1}{T}d(\boldsymbol{f}, \boldsymbol{p}_y)}}{\sum_{j=1}^{M} e^{-\frac{1}{T}d(\boldsymbol{f}, \boldsymbol{p}_j)}}, \tag{1}$$

$$\mathcal{L}_{cls}(\mathbb{D}^s) = -\frac{1}{N^s}\sum_{i=1}^{N^s} log(P(y_i^s|\boldsymbol{f}_i^s)). \tag{2}$$

From a perspective of probability, Eqn 1 can be viewed as the posterior probability of a data point belonging to a specific class with a mixture of exponential distribution where the prototypes act as the mean representations for each class [41]. Minimizing Eqn. 2 increases the posterior probability for each data point belonging to its labeled class. Therefore, data points will cluster around the corresponding class prototypes in the embedding space which conforms the cluster assumption. We further propose a compact regularizer resembling the contrastive loss in [8], which minimizes the distances between data points and their class prototypes to make each cluster more compact as follows:

$$\mathcal{L}_{reg}(\mathbb{D}^s) = \frac{1}{N^s}\sum_{i=1}^{N^s} d(\boldsymbol{f}_i^s, \boldsymbol{p}_{y_i^s}). \tag{3}$$

For prediction, we denote $C(y, \boldsymbol{f}; \{\boldsymbol{p}_j\}_{j=1}^M) = P(y|\boldsymbol{f})$ as a distance-based classifier with the class prototypes as its parameters. For a given input, we first obtain its feature with the feature extractor and then we classify it with the category of its nearest prototype, which is also the category with the maximum predicted probability.

3.2 Unsupervised Noise Removal

We aim to remove noisy source data so that they will not adversely affect domain adaptation [16, 29]. With discriminative clustering

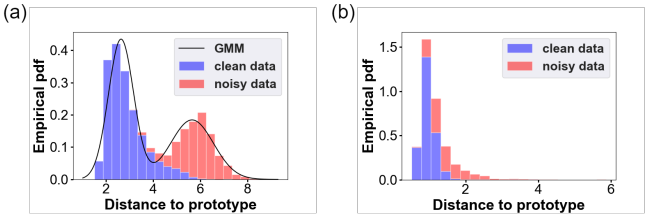

Figure 2: Empirical PDF and estimated GMM model on distance to class prototype for 40% mixed corruption in Amazon dataset after training a network for (a) 10 epochs and (b) 100 epochs.

from previous subsection, we observe that noisy source data locate quite differently in the embedding space compared to clean source data at the early training phase of deep networks. We train a deep network with objective $\mathcal{L}_{cls}+0.5*\mathcal{L}_{reg}$ for 100 epochs and measure the distribution of distances for both clean and noisy source data to their class prototypes at (a) epoch 10 (early phase) and (b) epoch 100 (late phase) as shown in Fig. 2. We observe that clean data locate closer to their class prototypes than almost all noisy data at the early phase of training but some noisy data also locate very close to their class prototypes at the late phase of training.

To understand this phenomenon, we refer to Arpit et al.'s study [1] on the memorization effect of deep networks. Arpit et al. find that when training with noisy data, deep networks learn simple patterns first before memorizing noisy data. Our observation conform Arpit et al.'s finding. At the early phase of training, as deep networks learn simple patterns first, clean data share simple patterns with each other will form class-wise clusters in the embedding space, thus clean data will locate closer to their class prototypes than noisy data. At the late phase of training, deep networks will memorize the complicated input-to-label patterns from label noise data and the complicated noise patterns from feature noise data, thus, noisy data will also locate close to their class prototypes.

The disparate distribution of distances to prototypes between clean and noisy source data at the early training phase of deep networks suggests the use of a two-component mixture model [3] to estimate the probability of a data point being clean based on its distance to the class prototype. In this paper, we propose a two-component Gaussian mixture model for this purpose as we empirically find it fits both the distance distribution of clean data and noisy data well as shown in Fig. 2(a). The probability density function of a two-component Gaussian distribution is defined as $p(d) = \sum_{k=1}^{2} \alpha_k p(d|k)$, where α_k is the prior probability for clean ($k = 1$) or noisy ($k = 2$) data and $p(d|k) = \mathcal{N}(d|\mu_k, \sigma_k)$ is the corresponding normal distance distribution with mean μ_k and covariance σ_k. We employ the Expectation-Maximization algorithm [6] to estimate the parameters of the Gaussian mixture model and calculate the posterior probability of a data point being clean as follows:

$$p(k=1|d) = \frac{\alpha_1 \mathcal{N}(d|\mu_1, \sigma_1)}{\sum_{k=1}^{2} \alpha_k \mathcal{N}(d|\mu_k, \sigma_k)}. \tag{4}$$

With the unsupervised Gaussian mixture noise model, we will remove data points with large probabilities of being noisy to prevent them affecting domain adaptation.

3.3 Cluster-Level Adversarial Adaptation

Noisy source data distort the real source data distribution, which make it more error-prone when aligning the data distribution across domains. More severely, due to label noise, the network cannot easily discriminate two data points from two different classes as they might be annotated with the same label. Thus there is no clear boundary between different classes in the embedding space and data from different classes would mix up with each other, which causes the class misalignment problem even more prominent for domain adaptation. Existing WSDA methods [29, 48] which align the marginal distribution across domains however would fail to resolve such issue and potentially transfer noise information from source domain to target domain. To address the problem, our idea is to align the target data with the more reliable class prototypes in the embedding space so that we can reduce the distribution shift across domains.

Specifically, we find that the prediction entropy of a data point encodes the location information of it in the embedding space. If a data point has low prediction entropy, it will locate close to some class prototypes in the embedding space (see Eqn. 1) and if a data point has high prediction entropy, it will locate near the decision boundaries between class prototypes. Source data cluster around their class prototypes in the embedding space, thus they will have low prediction entropy. However, due to the distribution shift across domains, except for some easy-to-transfer target data, which will locate close to some class prototypes in the embedding space like source data, other target data will locate near the decision boundaries [5, 26]. Thus, we design a domain discriminator based on a data point's prediction entropy. The domain discriminator is defined as $D(\boldsymbol{f}; \{p_j\}_{j=1}^{M}) = -\frac{1}{log(M)} \sum_{j=i}^{M} P(j|\boldsymbol{f}) log(P(j|\boldsymbol{f}))$, which calculates the normalized prediction entropy of a data point as the probability of the data point belonging to the target domain, where the $\frac{1}{log(M)}$ term is used to normalize the output within the interval $[0, 1]$.

Different from existing adversarial-based domain adaptation methods [10, 18, 32], our domain discriminator shares its parameters with the classifier. Since optimizing the classifier with the source data already ensures the prediction entropy on the source data is small, we train the domain discriminator with the cross entropy loss only on target data as follows:

$$\mathcal{L}_{adv_D}(\mathbb{D}^t) = -\frac{1}{N^t} \sum_{i=1}^{N^t} log(D(\boldsymbol{f}_i^t)). \tag{5}$$

Minimizing $\mathcal{L}_{adv_D}$ ensures the domain discriminator D can distinguish target data from source data based on their prediction entropy. To reduce the distribution shift across domains, we adversarially train the feature extractor with the GAN loss [11] as follows:

$$\mathcal{L}_{adv_F}(\mathbb{D}^t) = -\frac{1}{N^t} \sum_{i=1}^{N^t} log(1 - D(\boldsymbol{f}_i^t)). \tag{6}$$

Minimizing $\mathcal{L}_{adv_F}$ with the feature extractor will align target data towards their corresponding class prototypes in the embedding space to decrease their prediction entropy for confusing the domain discriminator. As class prototypes are the representative of each class in the embedding space, they are generally more reliable and

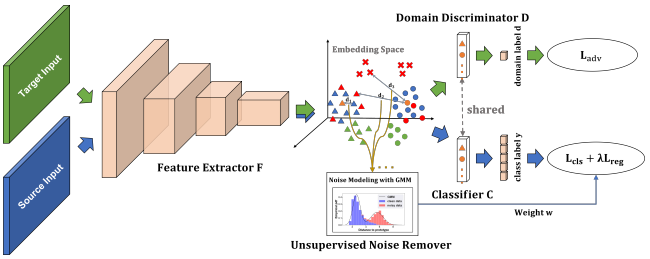

Figure 3: The overview of NTDA network architecture for WSDA with unsupervised noise removal and cluster-level adversarial adaptation. The network consists of a feature extractor, a domain discriminator, a label classifier and an unsupervised noise remover. Blue color represents the source domain, green color represents the target domain. In the embedding space, the red color represents noisy source data. Different shapes in the embedding space represents different classes and the "X" shape represents feature noise data.

less noisy compared to source data points, thus, aligning target data towards the less noisy class prototypes is beneficial in the WSDA scenario. In addition, our adaptation method maps target data towards the source data clusters at cluster level with consideration of source data's semantic structure, thus our method alleviates the class misalignment problem.

3.4 Optimization

In previous subsections, we have introduced different components of our network. In this subsection, we will combine these components into our final network and provide a simple and effective algorithm to train the network from end to end. First, to remove noisy source data, we propose a weighting scheme based on the Gaussian mixture noise model. Denote $d_i^s = \sqrt{d(\boldsymbol{f}_i^s, \boldsymbol{p}_{y_i^s})}$ as the euclidean distance of source data $\boldsymbol{x}_i^s$ to its class prototype $\boldsymbol{p}_{y_i^s}$ in the embedding space, the weight for $\boldsymbol{x}_i^s$ is defined as follows:

$$w(\boldsymbol{x}_i^s) = \mathbb{1}(p(1|d_i^s) > \eta)\frac{p(1|d_i^s) - \eta}{1 - \eta}, \tag{7}$$

where $\mathbb{1}(\cdot)$ is the indicator function, i.e., it returns 1 when the condition inside the brackets is true and returns 0 otherwise and η is the threshold hyper-parameter within interval $[0, 1]$. We weight the supervision loss for source data as follows:

$$\mathcal{L}_{cls_w}(\mathbb{D}^s) = -\frac{1}{N^s}\sum_{i=1}^{N^s} w(\boldsymbol{x}_i^s) log(P(y_i^s|\boldsymbol{f}_i^s)), \tag{8}$$

$$\mathcal{L}_{reg_w}(\mathbb{D}^s) = \frac{1}{N^s}\sum_{i=1}^{N^s} w(\boldsymbol{x}_i^s) d(\boldsymbol{f}_i^s, \boldsymbol{p}_{y_i^s}). \tag{9}$$

The weighting scheme only selects source data points whose probability of being clean is larger than η for training and it linearly scales the probability as the weight for the selected source data so that source data with higher probability of being clean will have larger weight. We present the network architecture of our model in Fig. 3 and the overall objective function to train the network is defined as follows:

$$\min_{\{\boldsymbol{p}_j\}_{j=1}^M} \mathcal{L}_{cls_w}(\mathbb{D}^s) + \lambda_1 \mathcal{L}_{reg_w}(\mathbb{D}^s) + \lambda_2 \mathcal{L}_{adv_D}(\mathbb{D}^t), \tag{10}$$

$$\min_{\Phi} \mathcal{L}_{cls_w}(\mathbb{D}^s) + \lambda_1 \mathcal{L}_{reg_w}(\mathbb{D}^s) + \lambda_2 \mathcal{L}_{adv_F}(\mathbb{D}^t), \tag{11}$$

where λ_1 and λ_2 are two trade-off hyper-parameters.

We present the algorithm to train the network in Alg. 1. Our algorithm warms up the network for N_p epochs to ensure the network learns some simple patterns first for unsupervised noise modeling. Note N_p can be chosen optimally via inspecting the distance distribution of source data as shown in Fig. 2. After warming up, our algorithm alternatively performs unsupervised noise removal and cluster-level adversarial adaptation. As our network trains mostly on clean source data, clean source data will be drawn closer and closer to the class prototypes in the embedding space which will further separate the distance distribution of clean and noisy source data apart and make our unsupervised noise model more accurate for selecting clean source data. This positive cycle between the unsupervised noise model and the network greatly boosts the performance of our method.

Algorithm 1: Training algorithm of NTDA model

Input: Warm up epoch N_p, Training epoch N_t, Feature extractor F, Classifier C, Domain discriminator D, Batch size B, T, η, λ_1, λ_2,

Output: Feature extractor F and Classifier C.

```
for N_p epoch do
    Train F and C on L_cls(D^s) + λ1 L_reg(D^s) with batch size B.
for N_t epoch do
    Model the distance distribution with a Gaussian mixture model on source data and calculates weights for source data with Eqn. 7.
    Train F, C and D on Eqn. 10 and Eqn. 11 with batch size B after removing source data with weight 0.
```

4 EXPERIMENTS

4.1 Datasets and Experimental Settings

Office-31 [25] is a benchmark dataset for domain adaptation, consisting of 4652 images with 31 classes in 3 distinct domains: Amazon (A), Webcam (W), DSLR (D). By permuting the 3 domains, we can generate 6 different domain adaptation tasks. **Office-Home** [34] is a more challenging dataset for visual domain adaptation, consisting of 15,500 images from 65 classes in 4 domains: Artistic (Ar), Clipart (Cl), Product (Pr) and Real-World (Rw). Similarly, we can generate 12 different domain adaptation tasks by permuting the 4 domains. **COVID-19** [47] is a cross-domain medical image analysis dataset for the diagnosis of COVID-19, which consists of 11663 images from 3 classes in two domains. The source domain consists of normal cases and pneumonia cases, while the target domain consists of normal cases and COVID-19 cases. We follow the studies in [47, 48] to transfer knowledge in identifying pneumonia cases to identify COVID-19 cases.

Method	Label Corruption							Feature Corruption							Mixed Corruption						
	A→W	W→A	A→D	D→A	W→D	D→W	Avg	A→W	W→A	A→D	D→A	W→D	D→W	Avg	A→W	W→A	A→D	D→A	W→D	D→W	Avg
ResNet [13]	47.2	33.0	47.1	31.0	68.0	58.8	47.5	70.2	55.1	73.0	55.0	94.5	87.2	72.5	58.8	39.1	69.3	37.7	75.2	75.5	59.3
DANN [10]	61.2	46.2	57.4	42.4	74.5	62.0	57.3	71.3	54.1	69.0	54.1	84.5	84.6	69.6	69.7	50.0	69.5	49.1	80.1	79.7	66.4
DAN [17]	63.2	39.0	58.0	36.7	71.6	61.6	55.0	73.9	60.2	72.2	59.6	92.5	88.0	74.4	64.4	45.1	71.2	44.7	79.3	78.3	63.8
RTN [19]	64.6	56.2	76.1	49.0	82.7	71.7	66.7	81.0	64.6	81.3	62.3	95.2	91.0	79.2	76.7	56.9	84.1	56.4	93.0	86.7	75.6
ADDA [32]	61.5	49.2	61.2	45.5	74.7	65.1	59.5	76.8	62.0	79.8	60.1	93.7	89.3	77.0	69.7	54.5	72.4	56.0	87.5	85.5	70.9
CDAN+E [18]	78.1	60.0	75.5	50.6	85.1	77.0	71.1	85.2	61.1	84.9	60.0	96.8	92.8	80.1	85.9	60.7	85.1	67.8	95.8	93.3	81.4
TCL [29]	82.0	65.7	83.3	60.5	90.8	77.2	76.6	84.9	62.3	83.7	64.0	93.4	91.3	79.9	87.4	64.6	83.1	62.2	**99.0**	92.7	81.5
CoUDA [48]	85.8	63.2	85.9	59.4	87.6	80.3	77.0	84.7	64.2	81.2	62.5	96.7	**93.3**	80.4	88.3	63.4	**89.7**	61.4	95.5	91.9	81.7
NTDA	**86.1**	**66.9**	**87.7**	**67.5**	**99.3**	**95.7**	**83.8**	**86.0**	**66.7**	**87.8**	**66.6**	**96.9**	93.0	**82.8**	**89.5**	**68.3**	87.1	**68.8**	98.7	**95.1**	**84.6**

Table 1: Classification Accuracy (%) on Office-31 with 40% corruption of labels and features.

Method	Office-Home													Bing-Caltech
	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg	B→C
ResNet [13]	27.1	50.7	61.7	41.1	53.8	56.3	40.9	28.0	61.8	51.3	33.0	65.9	47.6	74.4
DANN [10]	32.9	50.6	60.1	38.6	49.2	50.6	39.9	32.6	60.4	50.5	38.4	67.4	47.6	72.3
DAN [17]	40.9	54.2	63.0	47.2	54.3	56.3	47.2	42.8	69.0	61.0	47.4	71.9	54.6	75.0
RTN [19]	29.3	57.8	66.3	44.0	58.6	58.3	46.0	30.1	67.5	56.3	32.2	69.9	51.4	75.8
ADDA [32]	32.6	52.0	60.6	42.6	53.5	54.3	43.0	31.6	63.1	52.7	37.7	67.5	49.3	74.7
CDAN+E [18]	41.1	61.6	69.3	49.2	65.0	63.9	47.1	41.5	70.8	61.3	45.4	76.3	57.7	82.6
TCL [29]	38.8	62.1	69.4	46.5	58.5	59.8	51.3	39.9	72.3	63.4	43.5	74.0	56.6	79.0
CoUDA [48]	39.7	62.3	68.8	52.5	61.4	63.3	47.9	42.6	70.8	63.6	49.9	75.1	58.1	79.1
NTDA	**48.3**	**67.2**	**73.9**	**55.0**	**65.8**	**64.8**	**57.8**	**48.9**	**76.6**	**68.1**	**54.0**	**79.3**	**63.3**	**83.9**

Table 2: Classification Accuracy (%) on Office-Home with 40% mixed corruption and Bing-Caltech with native noise.

Since these three datasets are clean, we create their corrupted versions following the protocol in [29, 48]. In particular, we create noisy source data from the original clean data in three different ways: label corruption, feature corruption and mixed corruption. For label corruption, we change the label of each image uniformly to a random class with probability p_{noise}. For feature corruption, each image is corrupted by Gaussian blur and Salt-and-Pepper noise with probability p_{noise}. As for mixed corruption, each image is processed by label corruption and feature corruption with probability $p_{noise}/2$ independently. We term p_{noise} as the noise level for a domain adaptation task. In all the experiments, we use the noisy data for the source domain and clean data for the target domain.

Bing-Caltech [2] dataset contains Bing dataset and Caltech-256 dataset. The Bing dataset consists of images retrieved by Bing image search for each of the Caltech-256 category. Apart from the statistical differences between Bing images and Caltech images, the Bing dataset contains rich noise, with multiple objects in the same image. We use the Bing dataset as the noisy source domain and Caltech-256 as the clean target domain. While the experiments on Office-31, Office-Home and COVID-19 use manually synthesised noise, the experiments on Bing-Caltech report the performance for the real world weakly supervised domain adaptation.

Implementation Details. We adopt the 50-layer ResNet [13] as the feature extractor for all experiments on general images and MobileNet-V2 [27] as the feature extractor for all experiments on medical images. The hyper-parameter T is set to 10, λ_1 is set to 0.5, λ_2 is set to 1, η is set to 0.5. We employ SGD with weight decay 5e-4 to train the network. All experiments are repeated three times and we report the average result. For fair comparison, we report baseline results directly from the original papers if the experiment setting is the same and re-implement the methods to follow our experiment setting when there is difference.

Baseline Methods. We compare our method with state-of-the-art deep learning methods, domain adaptation methods, and weakly supervised domain adaptation methods. (1) **ResNet-50** [13] applies ResNet-50 trained on the noisy source domain to classify target data. (2) **MobileNet-V2** [27] applies MobileNet-V2 trained on the noisy source domain to classify the target data. (3) **DANN** [10] learns domain invariant features adversarially with gradient reversal. (4) **DAN** [17] applies multiple variants of MMD to align feature representations from multiple layers. (5) **RTN** [19] extends DAN by adapting classifiers through a residual transfer module. (6) **ADDA** [32] adopts asymmetric feature extractors for adversarial training. (7) **CDAN+E** [18] conditions the adversarial adaptation models on discriminative information conveyed in the classifier predictions and uses the entropy of prediction as an importance weight. (8) **CLAN** [20] conducts category-level domain adaptation. (9) **TCL** [29] learns a transferable curriculum for WSDA. (10) **CoUDA** [48] performs collaborative unsupervised domain adaptation for WSDA.

4.2 Experimental Analysis

Performance Comparison on General Images. We present the results on Office-31 under 40% label corruption, feature corruption and mixed corruption in Table 1 and the results on Office-Home under 40% mixed corruption and Bing-Caltech in Table 2. Our method outperforms all the baseline methods in almost all tasks and significantly pushes forward the state-of-the-art performance on Office-31 by improving the average accuracy by 6.8%, 2.4%, 2.9% on the label corruption, feature corruption and mixed corruption tasks respectively compared to the second best. It improves the average accuracy on Office-Home by 5.2%, and the accuracy on Bing-Caltech by 1.3% compared to the second best. For some tasks, the improvement of accuracy is more than 10%. More specifically, NTDA performs better than state-of-the-art WSDA methods TCL [29] and CoUDA [48],

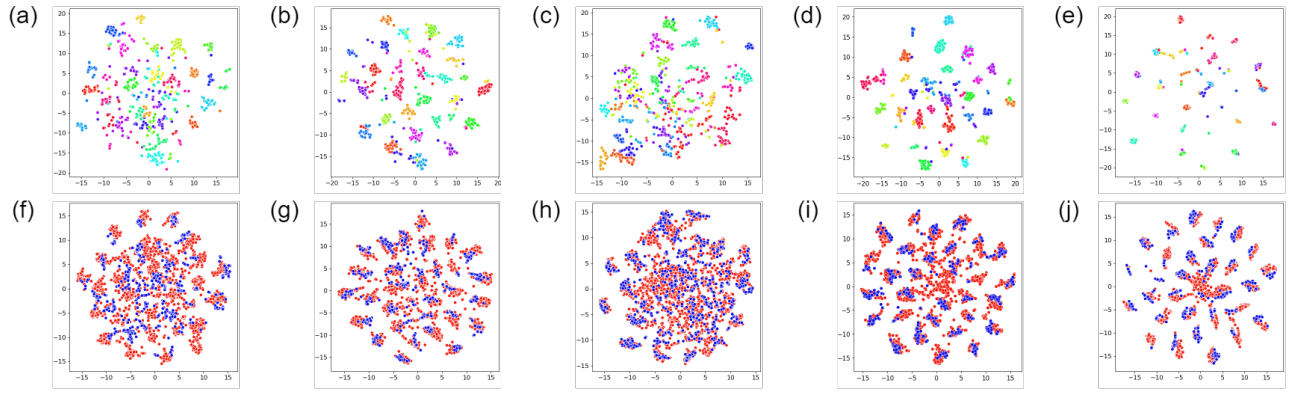

Figure 4: The t-SNE visualization of DANN, CDAN+E, TCL, CoUDA and NTDA with class labels (a)-(e) for target data and domain labels (f)-(j) for both source and target data in the embedding space. In (a)-(e), different colors represents different classes. In (f)-(j), red color represents source data and blue color represents target data.

indicating the effectiveness of our method for transferring knowledge from noisy source domain to unlabeled target domain. NTDA also outperforms state-of-the-art UDA methods by a large margin, indicating that existing UDA methods indeed suffer from the noise in source domain, thus it is necessary to come up with methods to counter the negative effects of noisy data.

Method	Acc (%)	MP	MR	Macro F1
MobileNet-V2 [27]	58.5	47.8	31.3	37.1
DANN [10]	87.2	51.6	55.6	51.1
CLAN [20]	95.6	75.0	71.1	72.9
TCL [29]	97.2	86.9	78.6	82.2
CoUDA [48]	98.1	94.3	82.4	87.3
NTDA	**98.3**	**99.3**	**89.2**	**93.6**

Table 3: Comparison on COVID-19 diagnosis with 10% label corruption.

Performance Comparison on Medical Images. Following [48], we apply 10% of label corruption on source domain of COVID-19 dataset and present the results in Table 3. We use Accuracy (Acc), Macro Precision (MP), Macro Recall (MP) and Macro F1-measure as metrics. Our method significantly outperform baseline methods in all metrics. More specifically, our method, NTDA, outperforms CoUDA [48], which is state-of-the-art method for the task. The results demonstrate the applicability of our method for cross-domain medical image analysis. In viewing that medical annotations are subjective and contain noises for domain adaptation, NTDA offers an appealing solution to the problem.

Method	A→W	W→A	A→D	D→A	W→D	D→W	Avg
NTDA (W/o UNR, CAA)	57.1	46.5	64.1	51.0	89.2	77.4	64.2
NTDA (W/o CAA)	74.4	62.9	78.5	57.8	97.4	90.9	74.2
NTDA (W/o UNR)	84.7	64.9	**88.3**	67.4	97.6	92.2	82.5
NTDA	**89.5**	**68.3**	87.1	**68.8**	**98.7**	**95.1**	**84.6**

Table 4: Classification Accuracy (%) on Office-31 with 40% mixed corruption.

Ablation Study. Table 4 shows the performance of different variants of our model on Office-31 with 40% mixed corruption. NTDA (W/o UNR, CAA) is the variant without Unsupervised Noise Removal and Cluster-Level Adversarial Adaptation. NTDA (W/o CAA) is the variant without Cluster-Level Adversarial Adaptation. NTDA (W/o UNR) is the variant without Unsupervised Noise Removal. Removing noisy source data and reducing the distribution shift across domains can both boost the target performance significantly, improving the accuracy by 10%, 18.3% respectively. NTDA combines unsupervised noise removal and cluster-level adversarial adaptation to achieve the best performance. Note NTDA (W/o UNR) also shows quite good performance, which is better than state-of-the-art WSDA methods, TCL [29] and CoUDA [48] on the same task. This is because our CAA method by itself is also robust to source data noise to some extent.

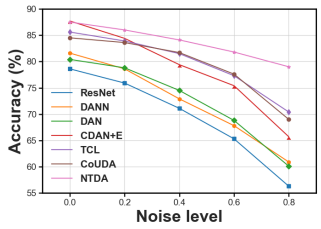

Figure 5: Classification Accuracy (%) w.r.t noise level.

Noisy Level. Fig. 5 shows the average classification results on Office-31 with mixed corruption under various noise levels. With the increase of noise level, the performance of all comparison methods degrade rapidly, while NTDA is more stable and provides much better performance which indicates that our method can handle various scenarios under weakly-supervised domain adaptation. In addition, NTDA performs as well as state-of-the-art domain adaptation method CDAN+E [18] even when noise level is 0, indicating that our method is also applicable in standard domain adaptation scenario.

Unsupervised Noise Modeling Quality. To test how well our unsupervised noise modeling method selects clean data, we compare it against the transferable curriculum in [29]. We use precision and recall as metrics for the comparison. Precision measures the fraction of clean data among the selected data, while recall measures the fraction of total number of clean data that are actually

Method		A→W	W→A	A→D	D→A	W→D	D→W	Avg
P(%)	TCL [29]	96.8	86.3	96.8	82.4	86.5	84.3	88.8
	NTDA	**99.2**	**96.2**	**98.2**	**95.3**	**96.3**	**96.6**	**97.0**
R(%)	TCL [29]	84.4	95.9	83.5	97.3	96.4	97.2	92.5
	NTDA	**92.3**	**98.2**	**94.3**	**99.6**	**98.2**	**100.0**	**97.1**

Table 5: Precision (P) and Recall (R) on Office-31 with 40% mixed corruption.

selected. Table 5 presents the precision and recall values of transferable curriculum and our method on Office-31 with 40% mixed corruption. In all tasks, NTDA provide larger precision and recall results than transferable curriculum. On average, NTDA can select 97.1% of clean data for training the network, and among the selected data only 3% are noisy data which are significantly better than transferable curriculum which can only select 92.5% of clean data, and among the selected data there are 11.2% noisy data. Our method can select almost all clean data and remove all noisy data for training.

Feature Visualization. Fig. 4 presents the t-SNE visualization of feature embeddings of DANN, CDAN+E, TCL, CoUDA and NTDA on task A→W with 40% mixed corruption. Fig. 4 (a)-(e) show the target feature embeddings by classes. NTDA model's embeddings are more compact and discriminative, while the rests' embeddings scatter and mix up among classes. Fig. 4 (f)-(j) show both the source and target feature embeddings by domain. NTDA aligns the target feature very well with the source data, while the other methods align the target feature not so well and mismatch the decision boundaries between the two domains. For DANN and CDAN+E, noisy source data degrades their feature embeddings. For TCL and CoUDA, as they align marginal distribution across domains without considering the fine-grained semantic structure in the embedding space, their feature embeddings surfer from class misalignment problem. These results validate the effectiveness of our method.

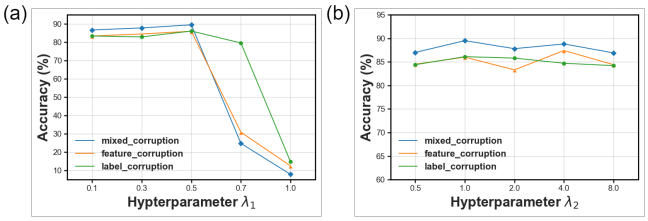

Figure 6: Classification Accuracy (%) on $A \rightarrow W$ under 40% mixed corruption w.r.t (a) λ_1 and (b) λ_2.

Hyper-parameters Sensitivity. Fig. 6(a) shows the sensitivity analysis of NTDA on hyper-parameter λ_1 under 40% label corruption, feature corruption and mixed corruptions. In general, NTDA performs stably when λ_1 is small, i.e. smaller than 0.5. When λ_1 becomes larger than 0.5, the performance of our method degrades. This is because λ_1 controls the strength of the compact regularizer to minimize the distances between data points and their class prototypes, when λ_1 becomes too large, it will tend to map all data points and class prototypes into a single point in the embedding space, thus making the performance much worse. Fig. 6(b) shows the sensitivity analysis of NTDA on hyper-parameter λ_2 under 40% label corruption, feature corruption and mixed corruption. In general, NTDA performs stably with the changes of λ_2. When λ_2 is within the interval $[0.5, 8]$, for label corruption and mixed corruption, the change of accuracy values is within 2% and for feature corruption, the change of accuracy is within 5%.

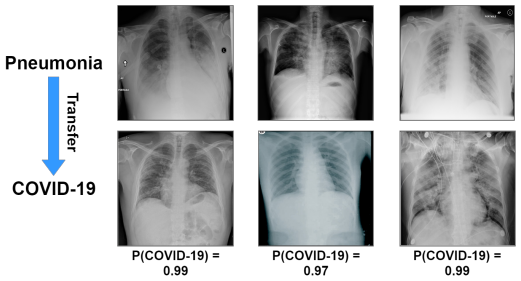

Figure 7: Samples analysis of NTDA for cross-domain COVID-19 diagnosis under 10% label corruption.

Sample Analysis. Fig. 7 shows the sample analysis of NTDA for the task of cross-domain COVID-19 diagnosis under 10% label corruption. Although the task of identifying pneumonia cases is different from the task of identifying COVID-19 cases. Pneumonia cases share some similar characteristics with COVID-19 cases [48]. But there also exists distribution difference between the two datasets, e.g., image color difference and different artifacts displayed in the images. Despite these differences and the fact that the annotations in the source data are noisy, NTDA can still provide high prediction probabilities for target images on the correct class, which demonstrates the effectiveness of our method when applied for medical image analysis.

5 CONCLUSIONS

In this paper, we study the under-explored Weakly Supervised Domain Adaptation problem (WSDA) for multimedia analysis. WSDA is a promising research area in view of its benefit to significantly reduce the annotation cost for deep learning. We identify several issues of existing WSDA methods and propose NTDA, a novel and effective method with unsupervised noise removal and cluster-level adversarial adaptation to alleviate the adverse effect of noisy data during domain adaptation. We conduct extensive experimental evaluation using four public datasets covering both general image analysis and medical image analysis, and the results show that our new method significantly outperforms existing methods.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their constructive comments, NUS colleagues and Beng Chin Ooi for their comments and contributions. This research is supported by Singapore Ministry of Education Academic Research Fund Tier 3 under MOE's official grant number MOE2017-T3-1-007. Meihui Zhang's work is supported by the National Natural Science Foundation of China (62050099).

REFERENCES

[1] Devansh Arpit, Stanisław Jastrzundefinedbski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. 2017. A Closer Look at Memorization in Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70* (Sydney, NSW, Australia) *(ICML'17)*. JMLR.org, 233–242.

[2] Alessandro Bergamo and Lorenzo Torresani. 2010. Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. In *Advances in neural information processing systems*. 181–189.

[3] Christopher M Bishop. 2006. *Pattern recognition and machine learning*. springer.

[4] Olivier Chapelle and Alexander Zien. 2005. Semi-supervised classification by low density separation.. In *AISTATS*, Vol. 2005. Citeseer, 57–64.

[5] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. 2019. Progressive feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 627–636.

[6] Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* 39, 1 (1977), 1–22.

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

[8] Zhijie Deng, Yucen Luo, and Jun Zhu. 2019. Cluster alignment with a teacher for unsupervised domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*. 9944–9953.

[9] Thomas Forgione, Axel Carlier, Géraldine Morin, Wei Tsang Ooi, Vincent Charvillat, and Praveen Kumar Yadav. 2018. An Implementation of a DASH Client for Browsing Networked Virtual Environment. In *Proceedings of the 26th ACM international conference on Multimedia*. 1263–1264.

[10] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*. PMLR, 1180–1189.

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

[12] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in neural information processing systems*. 8527–8537.

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

[14] Jonathan Krause, Benjamin Sapp, Andrew Howard, Howard Zhou, Alexander Toshev, Tom Duerig, James Philbin, and Li Fei-Fei. 2016. The unreasonable effectiveness of noisy data for fine-grained recognition. In *European Conference on Computer Vision*. Springer, 301–320.

[15] Lizi Liao, Xiangnan He, Bo Zhao, Chong-Wah Ngo, and Tat-Seng Chua. 2018. Interpretable multimodal retrieval for fashion products. In *Proceedings of the 26th ACM international conference on Multimedia*. 1571–1579.

[16] Feng Liu, Jie Lu, Bo Han, Gang Niu, Guangquan Zhang, and Masashi Sugiyama. 2019. Butterfly: A panacea for all difficulties in wildly unsupervised domain adaptation. In *NeurIPS LTS Workshop*.

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

[18] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. 2018. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*. 1640–1650.

[19] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. 2016. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*. 136–144.

[20] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. 2019. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2507–2516.

[21] Zhaojing Luo, Sai Ho Yeung, Meihui Zhang, Kaiping Zheng, Lei Zhu, Gang Chen, Feiyi Fan, Qian Lin, Kee Yuan Ngiam, and Beng Chin Ooi. 2021. MLCask: Efficient Management of Component Evolution in Collaborative Data Analytics Pipelines. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 1655–1666.

[22] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence* 41, 8 (2018), 1979–1993.

[23] Beng Chin Ooi, Kian-Lee Tan, Sheng Wang, Wei Wang, Qingchao Cai, Gang Chen, Jinyang Gao, Zhaojing Luo, Anthony KH Tung, Yuan Wang, et al. 2015. SINGA: A distributed deep learning platform. In *Proceedings of the 23rd ACM international conference on Multimedia*. 685–688.

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

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

[26] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. 2018. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3723–3732.

[27] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4510–4520.

[28] Rui Shu, Hung H. Bui, Hirokazu Narui, and Stefano Ermon. 2018. A DIRT-T Approach to Unsupervised Domain Adaptation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=H1q-TM-AW

[29] Yang Shu, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. 2019. Transferable Curriculum for Weakly-Supervised Domain Adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[30] Baochen Sun and Kate Saenko. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*. Springer, 443–450.

[31] Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoffrey J Gordon. 2020. Domain adaptation with conditional distribution matching and generalized label shift. *Advances in Neural Information Processing Systems* 33 (2020).

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

[33] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474* (2014).

[34] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. 2017. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5018–5027.

[35] Vedran Vukotić, Christian Raymond, and Guillaume Gravier. 2016. Multimodal and crossmodal representation learning from textual and visual features with bidirectional deep neural networks for video hyperlinking. In *Proceedings of the 2016 ACM workshop on Vision and Language Integration Meets Multimedia Fusion*. 37–44.

[36] Haoran Wang, Tong Shen, Wei Zhang, Ling-Yu Duan, and Tao Mei. 2020. Classes Matter: A Fine-grained Adversarial Approach to Cross-domain Semantic Segmentation. In *European Conference on Computer Vision*. Springer, 642–659.

[37] Wei Wang, Beng Chin Ooi, Xiaoyan Yang, Dongxiang Zhang, and Yueting Zhuang. 2014. Effective multi-modal retrieval based on stacked auto-encoders. *Proceedings of the VLDB Endowment* 7, 8 (2014), 649–660.

[38] Wei Wang, Xiaoyan Yang, Beng Chin Ooi, Dongxiang Zhang, and Yueting Zhuang. 2016. Effective deep learning-based multi-modal retrieval. *The VLDB Journal* 25, 1 (2016), 79–101.

[39] Yiling Wu, Shuhui Wang, Guoli Song, and Qingming Huang. 2019. Online asymmetric metric learning with multi-layer similarity aggregation for cross-modal retrieval. *IEEE Transactions on Image Processing* 28, 9 (2019), 4299–4312.

[40] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. 2015. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2691–2699.

[41] Hong-Ming Yang, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. 2018. Robust classification with convolutional prototype learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3474–3482.

[42] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. 2019. How does disagreement help generalization against label corruption?. In *International Conference on Machine Learning*. PMLR, 7164–7173.

[43] Xiyu Yu, Tongliang Liu, Mingming Gong, Kun Zhang, Kayhan Batmanghelich, and Dacheng Tao. 2020. Label-noise robust domain adaptation. In *International Conference on Machine Learning*. PMLR, 10913–10924.

[44] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. 2017. Central moment discrepancy (cmd) for domain-invariant representation learning. *arXiv preprint arXiv:1702.08811* (2017).

[45] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=Sy8gdB9xx

[46] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. 2021. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition. 12414–12424.

[47] Yifan Zhang, Shuaicheng Niu, Zhen Qiu, Ying Wei, P. Zhao, Jianhua Yao, Junzhou Huang, Qingyao Wu, and Mingkui Tan. 2020. COVID-DA: Deep Domain Adaptation from Typical Pneumonia to COVID-19. *ArXiv* abs/2005.01577 (2020).

[48] Yifan Zhang, Ying Wei, Qingyao Wu, Peilin Zhao, Shuaicheng Niu, Junzhou Huang, and Mingkui Tan. 2020. Collaborative unsupervised domain adaptation for medical image diagnosis. *IEEE Transactions on Image Processing* 29 (2020), 7834–7844.