

Hierarchical Unsupervised Relation Distillation for Source Free Domain Adaptation

Bowei Xing, Xianghua Ying*, Ruibin Wang, Ruohao Guo,
Ji Shi, and Wenzhen Yue

National Key Laboratory of General Artificial Intelligence, School of Intelligence
Science and Technology, Peking University

{xingbowei,xhying,robin_wang, sjj118}@pku.edu.cn
ruohguo@foxmail.com, yuewenzhen@stu.pku.edu.cn

Abstract. Source free domain adaptation (SFDA) aims to transfer the model trained on labeled source domain to unlabeled target domain without accessing source data. Recent SFDA methods predominantly rely on self-training, which supervise the model with pseudo labels generated from individual data samples. However, they often ignore the crucial data structure information and sample relationships that are beneficial for adaptive training. In this paper, we propose a novel hierarchical relation distillation framework, establishing multi-level relations across samples in an unsupervised manner, which fully exploits inherent data structure to guide sample training instead of using isolated pseudo labels. We first distinguish source-like samples based on prediction reliability during training, followed by an effort on distilling knowledge to those target-specific ones by transferring both local clustering relation and global semantic relation. Specifically, we leverage the affinity with nearest neighborhood samples for local relation and consider the similarity to category-wise Gaussian Mixtures for global relation, offering complementary supervision to facilitate student learning. To validate our approach’s effectiveness, we conduct extensive experiments on diverse benchmarks, achieving better performance compared to previous methods.

Keywords: Source Free Domain Adaptation · Relation Distillation · Knowledge Transfer

1 Introduction

Deep neural network (DNN) models have achieved remarkable success in various visual recognition tasks. However, even very large DNN models often suffer significant performance degradation when there is a distribution or domain shift [28] between training (source) and testing (target) domains. To address the problem of domain shifts, various Unsupervised Domain Adaptation (UDA) algorithms [1, 5, 22, 31] have been developed, which aim to transfer the knowledge from labeled source domain to the unlabeled target domain. Due to data

* Corresponding author.

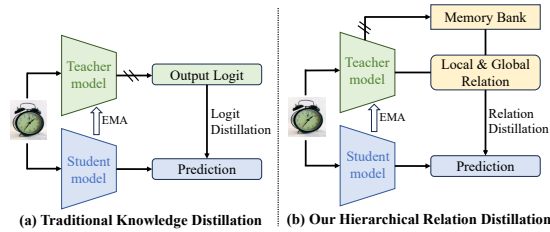


Fig. 1: (a) Traditional knowledge distillation transfers the individual output logit of isolated sample from teacher to student. (b) Our relation distillation integrates structural relations of overall dataset as supervision for student learning.

privacy concerns, more recent methods focus on Source Free Domain Adaptation (SFDA) [4, 17, 40], which is a more practical setting. Under this setting, the labeled source data is not accessible any more when training the target model, but only the pre-trained model from the source domain is provided.

Recently, several SFDA methods [12, 15, 18–20, 37, 38, 43, 45] have been proposed, focusing on mutual information maximization [19], self-training [14, 20], auxiliary self-supervised tasks [2, 3, 45] and so on. Among them, self-training methods have emerged as a promising solution for SFDA, which generate pseudo labels to supervise model training for target data. However, the performance of self-training can suffer from noise due to domain shifts. Since the pseudo labels are always generated from the model prediction of individual sample, lacking consideration of intrinsic data structure, the optimization of the source-trained model is then degraded to entropy minimization [18], where the updated model can achieve high confident but wrong predictions on account of pseudo label noise. Several works [12, 35, 45] propose to divide the samples into different sets and process them separately, however, they ignore the relation between sample sets, missing to exploit the reliable information from those confident samples.

In this paper, we identify the above limitations and propose a novel hierarchical relation distillation (HRD) framework that rectifies feature learning by harnessing the intrinsic data structure, in the meanwhile exploiting different properties including local and global relations across samples with varying reliability. Specifically, we first divide the target samples into two pseudo domains, named as the source-like set and target-specific set, employing a class-balanced division strategy that accentuates the former’s heightened prediction reliability with the source model. Given the absence of accessible ground truth labels, we intend to construct relations between the two pseudo domains unsupervisedly, and then distill the reliable structural knowledge to facilitate the sample training and model adaptation with a teacher-student scheme. In order to enable the modeling of longer range and broader structure information, we introduce a memory bank to store the output features and logits of target data, enabling hierarchical relationship establishment spanning both local and global contexts. Finally, the two relation distillation branches are merged through an adaptive

ensemble strategy, regularized by a reliability-aware weighting coefficient, thus to fully utilize the information from different properties.

The core idea lies in exploiting the inherent relations across samples to facilitate sample training in a more robust manner and transfer knowledge from those well-learned samples to the other ones, *i.e.*, from source-like set to target-specific set. Specifically, we introduce two pivotal constructs: local clustering relation and global semantic relation. Since visually similar samples are closer than dissimilar ones in feature space, we propose to leverage the information of local context to facilitate the cultivation of sample learning. Concretely, we calculate local affinity with the top-K nearest source-like neighbors and regard these reliable neighbor predictions as soft labels, leading to the local clustering relation, which also enforces aligned representations in the neighborhood. As for the global semantic relation, we dynamically build Gaussian Mixtures to represent the distribution of each specific category, utilizing source-like samples as the centroids of different Gaussian mixtures. On this basis, it enables us to establish semantic relation by measuring the relative similarity to each class-wise mixture for the given sample, thus to facilitate model adaptation from a global probabilistic perspective. Overall, the former relation focuses on local spatial alignment and the latter attends to global semantic distribution, offering complementary supervision to improve sample training.

Finally, we conduct comprehensive experiments on several popular SFDA benchmarks, including Office-31 [29], Office-Home [33], VisDA-2017 [26] and DomainNet [25]. The results demonstrate that the proposed HRD achieves superior performance against comparable state-of-the-art SFDA methods, in the meanwhile exhibiting compatibility with the recent strong baselines [19, 38]. Additionally, the ablation studies validate the effectiveness of individual components of our framework for model adaptation. In summary, our contributions are:

1. We propose a novel perspective for source free domain adaptation from the view of relation distillation, exploiting data structure information and sample relationships to facilitate the adaptation in target domain.
2. We establish the hierarchical relations among samples unsupervisedly, proposing local clustering relation and global semantic relation. We also employ an effective adaptive ensemble strategy to regularize these two branches.
3. The experimental results demonstrate that our approach outperforms existing SFDA methods on several datasets.

2 Related Work

2.1 Source Free Domain Adaptation

SFDA aims to transfer the model from source domain to target domain without accessing source data. Recently, several approaches [4, 12, 18–20, 24, 38–40] have been proposed for SFDA. Specifically, SHOT [19] utilizes a prototype-based method to obtain pseudo label for self-training. However, pseudo labels generated at the early training stage can be noisy, and several works propose

different strategies to refine the pseudo labels. GPUE [20] proposes to guide pseudo label learning by combining with uncertainty estimation. G-SFDA [37] and NRC [39] follow a similar strategy by encouraging consistent predictions between the neighborhood samples. In addition, DaC [45] considers sample division and takes different strategies for the two subsets, e.g. generating different contrastive positive pairs for each sample. Although great success has been made, the crucial target data structure and sample relationships are not fully exploited in these works. To this end, we propose to rectify label noise by integrating local and global data structure without any rigorous and expensive pseudo label refinement technique, in the meanwhile utilizing relation distillation to mine reliable information and bring benefits for the unreliable data learning.

2.2 Knowledge Distillation

Knowledge distillation [6, 9, 46] aims to transfer knowledge from a teacher model to a student model. By leveraging knowledge distillation, the student model could learn implicit knowledge by incorporating an extra objective of the teacher’s outputs so as to gain better performance. Recently, some works [41, 42] have demonstrated that distilling models of identical architecture in a teacher-student scheme, *i.e.*, self-distillation, can further improve the performance of neural network. Specifically, they use the semantically meaningful probability distribution produced by the teacher model to supervise the learning of the student model. However, since the outputs of the teacher model are only based on processing the isolated sample, the noise of the prediction logits may lead to wrong optimization direction. Therefore, we explore to consider a broader view by making use of data structure and sample relationships to obtain more appropriate probability distribution for unlabeled target samples, as depicted in Fig. 1. In this way, the distillation is no longer based on single instance, but based on the correlation of target samples, leading to more reliable and robust supervision for student.

3 Method

3.1 Preliminary

In the domain adaptation setting, a labeled source domain dataset and an unlabeled target domain dataset are provided, with the aim to learn an adaptive model transferred from source to target. Concretely, the source dataset $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ contains N_s labeled samples distributed over C classes. Here, $y_i^s \in \mathcal{Y} \subseteq \mathbb{R}^C$ denotes the one-hot ground-truth label for sample x_i^s . The model $\phi = f \circ g$ can be regarded as the combination of a feature extractor f and a classifier g , where $f : x \rightarrow \mathbb{R}^D$ maps the input image to a D -dimensional feature and $g : \mathbb{R}^D \rightarrow \mathbb{R}^C$ further maps it to a C -dimensional category probability vector.

In this paper, we focus on SFDA problem where the data in $\mathcal{D}_s$ is unavailable while only the pre-trained source model ϕ is available. The target domain dataset is denoted as $\mathcal{D}_t = \{(x_i^t)\}_{i=1}^{N_t}$ containing N_t unlabeled samples, which shares the same label set $\mathcal{Y}$ as $\mathcal{D}_s$.



The overall framework of our proposed method is illustrated in Fig. 2. Following the popular teacher-student paradigm, our proposed hierarchical relation distillation framework adopts a pair of teacher model ϕ_t and student model ϕ_s with exactly the same network architecture. In each training iteration, the original and weak-augmented target samples are fed to the teacher and student, respectively. Considering that the individual pseudo label directly and simply from the teacher model contains noise and may lead to wrong optimization directions, we intend to include more robust and reliable supervision from the teacher.

3.3 Hierarchical Relation Distillation

Class-Balanced Sample Division. To begin with, we will describe the process of dataset division, through which we intend to distinguish samples that are learnt well and not that well with source model. In general, the accuracy of sample can be strongly related to prediction confidence [7], therefore, we take advantage of the prediction confidence as a measurement. Intuitively, when

processed by source model, the samples that have confident predictions exhibit higher certainty and can be regarded as more similar to source domain distribution. In this way, we can divide the samples into two pseudo domains according to prediction reliability. Besides, the learning effect can also be various for different categories, therefore, the division principle should be related to different classes so as to reflect the class-level learning status. Moreover, the threshold should be increasing steadily as the model is continuously improved during training.

To this end, we propose a class-balanced division strategy based on the prediction confidence, in the meanwhile implemented with an adaptive threshold to be fit with different learning stages and different categories. Concretely, the samples are dynamically grouped into the source-like set by threshold τ_c for the c -th category, and then forming the source-like set $\mathcal{D}_r$ and target-specific set $\mathcal{D}_u$:

$$\begin{aligned}\mathcal{D}_r(c) &= \{x_i | \max(p_i) > \tau_c, x_i \in \mathcal{O}(c)\} \\ \mathcal{D}_r &= \cup_{i=1}^C \mathcal{D}_r(c), \quad \mathcal{D}_u = \mathcal{D}_t / \mathcal{D}_r\end{aligned}\tag{1}$$

where p_i is a C -dimensional vector from the teacher classifier output, $\mathcal{O}(c) = \{x_i | \operatorname{argmax}(p_i) = c, x_i \in \mathcal{D}_t\}$ contains samples that have the highest prediction score on class c , and $\tau_c = \frac{1}{|\mathcal{O}(c)|} \sum_{x_i \in \mathcal{O}(c)} p_i^c$ is a class-balanced threshold value to filter samples without fine-tuning.

Local Clustering Relation. After feature extraction, similar samples are more likely to have similar features and lead to clustering in the feature space, therefore, we propose to incorporate information of the neighboring target samples to estimate the local distribution that are more robust against domain shifts. The local distribution can be obtained by considering the relation with reliable neighbors for the anchor sample, so as to transfer the local structural information from teacher to student.

Specifically, given anchor sample x_i , we first calculate its local affinity to the reliable neighborhood samples. Suppose that K nearest samples for x_i are selected from the source-like set, we can obtain a K -dimensional affinity vector based on the cosine distance between the anchor sample and the neighborhood samples. The affinity vector reflects the similarity among other samples for x_i in the local region, where the more similar sample is assigned higher affinity and vice versa. In formulation, this K -dimensional normalized affinity vector can be written as:

$$A_i = [a_{i,\mathcal{N}(i)}], \quad a_{i,j} = \frac{\cos(f(x_i), f(x_j))}{\sum_{k \in \mathcal{N}(i)} \cos(f(x_i), f(x_k))}\tag{2}$$

where $\mathcal{N}(i) = \{j | x_j \in \text{KNN}(x_i, K, \mathcal{D}_r)\}$ represents the top- K nearest neighbor samples around x_i in $\mathcal{D}_r$.

Since the neighborhood samples are selected from source-like set and exhibit high confidence, we thus see them as trustworthy support samples and regard their prediction logits as soft class labels, leading to soft label vector $P_i = [p_{n(1)}, p_{n(2)}, \dots, p_{n(K)}]$, where we re-write $\mathcal{N}(i)$ as $\{n(1), n(2), \dots, n(K)\}$.

Therefore, the anchor sample can be aware of local category distribution by aggregating soft labels with local affinity, leading to local clustering relation:

$$r_i^{local} = A_i \cdot P_i^T \quad (3)$$

In this way, the obtained vector r_i^{local} contains the relationship with local reliable neighbors instead of based on single sample prediction, thus can be more robust and reliable. We can distill the aggregated relationships of each sample with respect to the local neighbors in target domain, which encourages the model to maintain more stable representations by anchoring a broader view.

Global Semantic Relation. In addition to the local clustering relation, we would like to further utilize the global data structure to establish the global relation, by considering semantic similarity with the data from the whole dataset. Concretely, we utilize Gaussian Mixture Model (GMM), a typical probabilistic model, to represent the distributions of different categories, which can be further introduced to formulate the global similarity. Since we have distinguished the well-learned samples that contain reliable semantic information, on this basis, we regard them as the centroids of Gaussian Mixtures and build a reliable GMM, enabling to model the relationship among samples and integrate the overall target structure information.

In order to build a GMM which includes C Gaussian mixture components, we firstly calculate the average features of the source-like samples belonging to c -th class as the mean μ_c with the reliable samples as centers:

$$\mu_c = \frac{1}{|\mathcal{D}_r(c)|} \sum_{x_i \in \mathcal{D}_r(c)} f(x_i) \quad (4)$$

where $f(x_i)$ is the extracted feature representation of sample x_i , $\mathcal{D}_r(c)$ is the reliable set for class c . After obtaining μ_c , the variance σ_c of c -th component can be calculated by measuring the distance between samples and mixture centers:

$$\sigma_c = \sqrt{\frac{\sum_{x_i \in \mathcal{D}_r} p_i^c \cdot (f(x_i) - \mu_c)^2}{\sum_{x_i \in \mathcal{D}_r} p_i^c}} \quad (5)$$

where p_i^c is the prediction score of the c -th category for x_i .

On this basis, we can utilize the established reliable GMM to model the global semantic similarity for the given sample. To be specific, by comparing each respective sample feature to the Gaussian Mixtures, we obtain the global semantic relation, which can be formulated as:

$$\begin{aligned} g(x_i | \mu_c, \sigma_c) &= \frac{1}{\sqrt{2\pi\sigma_c^2}} \cdot \exp\left(-\frac{(f(x_i) - \mu_c)^2}{2\sigma_c^2}\right) \\ p(x_i)_c &= \frac{\pi_c \cdot g(x_i | \mu_c, \sigma_c)}{\sum_{j=1}^C \{\pi_j \cdot g(x_i | \mu_j, \sigma_j)\}} \\ r_i^{global} &= [p(x_i)_1, p(x_i)_2, \dots, p(x_i)_C] \end{aligned} \quad (6)$$

where $\pi_j = \frac{|\mathcal{D}_r(j)|}{|\mathcal{D}_r|}$ is the mixing coefficient calculated in GMM, indicating the probability that the j -th component is selected. Note that, the employed GMM is implemented adaptively, where the parameters are dynamically adapted with the training going. As the features and predictions can be varied in each epoch, it introduces more randomness that further regularizes the model training. Besides, the number of reliable samples can be progressively increased, thus a more reasonable GMM can be constructed, gradually approximating the real target distribution and transferring more reliable global structural knowledge.

Adaptive Ensembled Distillation. Because the ground truth label is not available and the prediction of teacher model can not be exactly accurate, we additionally add an adaptive regularization for distillation. Concretely, we propose an adaptive weighting coefficient for each sample’s relation distillation process and also balance the two distillation branches, which exclusively regularizes the loss function without incurring additional computational overhead.

Specifically, we adopt the reliability of the predictions to measure the respective proportion of each local and global relation vector as self-regularization, then adaptively ensemble them for the final optimization. To measure the label reliability, we exploit the max score and the entropy of the relation distribution to assign adaptive weight for each branch and for each sample, using the product of these two metrics as the adaptive weight, written as:

$$\omega_i = \max(r_i) * \exp(-\text{entropy}(r_i)) \quad (7)$$

In this way, we can consider the reliable prediction more and the unreliable prediction less in an adaptive manner, achieving sample-wise re-weighting and branch balancing, thus to improve the distillation. Overall, the adaptively regularized hierarchical relation distillation objective is:

$$\begin{aligned} \mathcal{L}_{rd} = \mathbb{E}_{x_i \in \mathcal{D}_t} [& \omega_i^{local} \cdot KL(\sigma(\frac{r_i^{local}}{\rho}) || \sigma(\frac{\phi_s(x_i)}{\rho})) \\ & + \omega_i^{global} \cdot KL(\sigma(\frac{r_i^{global}}{\rho}) || \sigma(\frac{\phi_s(x_i)}{\rho}))] \end{aligned} \quad (8)$$

where σ denotes the softmax function, ρ is a pre-defined temperature, and $\phi_s(x_i)$ is the student prediction of x_i .

3.4 Overall Training Objective

For the model optimization, the student is updated by the loss propagation. Concretely, in addition to the above hierarchical relation distillation loss, we also include a widely used diversity loss in SFDA methods [19, 39]:

$$\mathcal{L}_{div} = \sum_{c=1}^C \hat{p}^c \log(\hat{p}^c) \quad (9)$$

where $\hat{p} = \mathbb{E}_{x_i \in \mathcal{D}_t}[\phi_s(x_i)]$ is the mean output embedding of the whole target domain samples from the student model.

Combining the above losses with a coefficient β as trade-off (setting $\beta = 1$ as previous works [19, 39]), the overall optimization objective is summarized as:

$$\mathcal{L} = \mathcal{L}_{rd} + \beta \mathcal{L}_{div} \quad (10)$$

For the teacher update, we simply take the exponential moving average of student parameters θ_s to update the teacher parameters θ_t during each iteration.

$$\theta_t = \lambda \theta_t + (1 - \lambda) \theta_s \quad (11)$$

where λ is an updating coefficient.

It is noted that our proposed relation distillation framework can also be combined with other methods, e.g. replacing the diversity loss with the dispersing loss in AaD [38], which is denoted as HRD++ in the experiments.

4 Experiments

4.1 Experimental Setup

Datasets. We carry out experiments on four benchmarks for image classification. Concretely, Office-Home [33] has a total of 15.5K images from 65 classes collected from 4 different image domains: Artistic, Clipart, Product, and Real-world, constituting 12 transfer tasks. Office-31 [29] is a small-scale benchmark with images from 31 categories across 3 domains: Amazon, DSLR and Webcam. VisDA [26] contains 2 different domains, synthetic and real, with 12 shared classes. We consider the model adaptation from synthetic to real domain, and report the average over per-class accuracy. For DomainNet, following previous works [45], we use a subset of it that contains 126 classes from 4 domains (Real, Sketch, Clipart, Painting) and evaluate 7 domain shifts built from them.

Implementation Details. For the sake of a fair and direct comparison, we use the same experimental setup and architecture as previous works. Specifically, we use ResNet50 [8] as backbone for Office-31, Office-Home, DomainNet and ResNet101 [8] for VisDA. For source model training, we initialize the ResNet backbone with ImageNet-1K pre-trained weights, and then the model is trained with the standard cross-entropy loss and with label-smoothing like in [19]. For the adaptation in target domain, the model is first initialized with the source model’s parameters. During training, we adopt SGD optimizer with a momentum of 0.9 and a weight decay of 1e-4. The learning rate is set as 5e-3 for Office datasets and 5e-4 for VisDA and DomainNet, while the batch size is set as 64 for all datasets. For the input samples of student model, we use weak data augmentation that consists of resized crops and horizontal flip. As for hyper-parameters, we set K as 3 for Office datasets and 5 for the other two datasets, the EMA update coefficient λ is set as 0.99 and the temperature ρ is set as 0.1 empirically.

Table 1: Accuracy (%) on Office-Home and VisDA. We highlight the **best** and the **second** Avg results for each benchmark. SF in the table is short for source-free.

Method	SF	Office-Home												VisDA	
		A→C	A→P	A→R	C→A	C→P	C→R	P→A	P→C	P→R	R→A	R→C	R→P	Avg	Avg
ResNet-50 [8]	✗	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1	52.4
FixBi [23]	✗	58.1	77.3	80.4	67.7	79.5	78.1	65.8	57.9	81.7	76.4	62.9	86.7	72.7	87.2
SCDA [16]	✗	60.7	76.4	82.8	69.8	77.5	78.4	68.9	59.0	82.7	74.9	61.8	84.5	73.1	84.2
SHOT [18]	✓	57.1	78.1	81.5	68.0	78.2	78.1	67.4	54.9	82.2	73.3	58.8	84.3	71.8	82.9
A ² Net [36]	✓	58.4	79.0	82.4	67.5	79.3	78.9	68.0	56.2	82.9	74.1	60.5	85.0	72.8	84.3
G-SFDA [37]	✓	57.9	78.6	81.0	66.7	77.2	77.2	65.6	56.0	82.2	72.0	57.8	83.4	71.3	85.4
NRC [39]	✓	57.7	80.3	82.0	68.1	79.8	78.6	65.3	56.4	83.0	71.0	58.6	85.6	72.2	85.9
AaD [38]	✓	59.3	79.3	82.1	68.9	79.8	79.5	67.2	57.4	83.1	72.1	58.5	85.4	72.7	88.0
DaC [45]	✓	59.1	79.5	81.2	69.3	78.9	79.2	67.4	56.4	82.4	74.0	61.4	84.4	72.8	87.3
Sub-Sup [13]	✓	61.0	80.4	82.5	69.1	79.9	79.5	69.1	57.8	82.7	74.5	65.1	86.4	74.0	88.2
BMD [27]	✓	58.1	79.7	82.6	69.3	81.0	80.7	70.8	57.6	83.6	74.0	60.0	85.9	73.6	88.7
CREL [43]	✓	62.8	82.0	84.3	70.9	80.8	82.6	70.0	61.1	83.6	76.2	65.1	87.0	75.5	89.1
UTR [24]	✓	59.8	81.2	83.2	67.2	79.2	80.1	68.4	56.4	83.0	73.7	61.2	85.9	73.2	87.5
C-SFDA [12]	✓	60.3	80.2	82.9	69.3	80.1	78.8	67.3	58.1	83.4	73.6	61.3	86.3	73.5	87.8
HRD (Ours)	✓	61.2	82.6	84.5	70.8	81.4	82.5	69.9	59.7	84.0	75.2	63.6	87.4	75.2	89.8
HRD++ (Ours)	✓	63.6	83.6	85.4	71.2	83.3	83.9	70.1	60.9	85.2	77.3	64.5	88.5	76.5	90.2

4.2 Comparison Results

We report Top-1 accuracy on the above benchmarks to validate the performance of HRD against previous methods. For Office-Home and Office-31, the accuracy for each domain shift and on average is shown in Tab. 1 and Tab. 2. Compared with the baseline method (*i.e.*, SHOT [19]), our HRD achieves an average 3.4% accuracy improvement on Office-Home and an average 2.0% improvement on Office-31, respectively. Since we use the same architecture and experimental setup as the baseline, the improvements can be directly attributed to our proposed novel hierarchical relation distillation framework. Besides, when combined with the dispersing loss as AaD [38], we achieve the state-of-the-art results on these two benchmarks. We believe that, exploiting the structural relation among samples and propagate the knowledge from source-like to target-specific ones is beneficial for the source model adaptation.

On the other two larger scale and more challenging datasets, *i.e.*, VisDA and DomainNet, we also achieve better performance comparing with both UDA and SFDA methods, as shown in Tab. 1 and Tab. 3. It is noted that, even without accessing source data during adaptation, we can outperform UDA methods [16, 23] by a large margin. Compared to the recent SFDA methods, it is clear that both HRD and HRD++ outperform previous approaches, where HRD++ achieves the best performance and HRD is the second best. In particular, compared to DaC [45] that also employs a data division strategy for SFDA but lacks consideration of relations between different sets, we significantly outperform it by 2.9% on VisDA and 2.0% on DomainNet, demonstrating the superiority of our proposed relation distillation. Besides, HRD achieves competitive performance in most kinds of domain shifts, validating its generality for different types of data. It is also worth mentioning that the success of HRD++ demonstrates the compatibility of our method when selecting AaD as a stronger baseline.

Table 2: Accuracy (%) on Office-31.

Method	SF	A→D	A→W	D→W	W→D	D→A	W→A	Avg
MDD [44]	✗	90.4	90.4	98.7	99.9	75.0	73.7	88.0
MCC [10]	✗	95.6	95.4	98.6	100.0	72.6	73.9	89.4
SRDC [32]	✗	95.8	95.7	99.2	100.0	76.7	77.1	90.8
SCDA [16]	✗	95.4	95.3	99.0	100.0	77.2	75.9	90.5
SHOT [18]	✓	94.0	90.1	98.4	99.9	74.7	74.3	88.6
3C-GAN [15]	✓	92.7	93.7	98.5	99.8	75.3	77.8	89.6
NRC [39]	✓	96.0	90.8	99.0	100.0	75.3	75.0	89.4
BMD [27]	✓	96.2	94.2	98.0	100.0	76.0	76.0	90.1
AaD [38]	✓	96.4	92.1	99.1	100.0	75.0	76.5	89.9
VMP [11]	✓	93.3	96.2	98.6	100.0	75.4	76.9	90.0
C-SFDA [12]	✓	96.2	93.9	99.7	98.8	77.9	77.3	90.5
UTR [24]	✓	95.0	93.5	99.9	99.1	78.4	76.3	90.3
CREL [43]	✓	95.8	95.1	99.0	100.0	76.6	78.3	90.8
HRD (Ours)	✓	96.3	94.1	98.9	99.1	77.4	77.6	<u>90.6</u>
HRD++ (Ours)	✓	96.5	94.5	99.1	99.9	77.1	77.9	90.8

Table 3: Accuracy (%) on DomainNet. The * baseline is implemented by official codes.

Method	SF	R→C	R→P	P→C	C→S	S→P	R→S	P→R	Avg
MCC [30]	✗	44.8	65.7	41.9	34.9	47.3	35.3	72.4	48.9
CDAN [21]	✗	65.0	64.9	63.7	53.1	63.4	54.5	73.2	62.5
SHOT [18]	✓	67.7	68.4	66.9	60.1	66.1	59.9	80.8	67.1
TENT [34]	✓	58.5	65.7	57.9	48.5	52.4	54.0	67.0	57.7
AdaCon [2]	✓	70.2	69.8	68.6	58.0	65.9	61.5	80.5	67.8
AaD* [38]	✓	70.5	69.9	69.8	60.6	65.7	61.7	78.3	68.1
DaC [45]	✓	70.0	68.8	70.9	62.4	66.8	60.3	78.6	68.3
GPUE [20]	✓	74.2	70.4	68.8	64.0	67.5	65.7	76.5	69.6
C-SFDA [12]	✓	70.8	71.1	68.5	62.1	67.4	62.7	80.4	69.0
HRD (Ours)	✓	71.7	71.2	70.3	64.5	68.2	62.7	80.8	<u>69.9</u>
HRD++ (Ours)	✓	71.9	71.6	71.1	64.7	68.5	62.5	81.8	70.3

4.3 Ablation Study

Effectiveness of hierarchical relation distillation. Tab. 4 gives the contribution of each key component that employed in our method. Specifically, we detail the performance of each relation distillation branch as well as its accuracy improvement by progressively adding them. It can be observed that the two modules efficiently capture structure information and consistently improve the performance over the baseline. Either considering local or global relation can significantly outperforms the conventional pseudo label that based on individual sample. By combining both, it achieves the highest accuracy, verifying the complementarity of the employed hierarchical relation scheme.

Effectiveness of adaptive ensembled strategy. In our method, we employ the reliability to re-weight the two distillation branches and assign adaptive regularization for each sample. On the contrary, we also implement the framework without the re-weighting coefficient, *i.e.*, by directly adding the two KL losses. As shown in Tab. 4, when combined with the adaptive re-weighting strategy, a consistent improvement is achieved on all four benchmarks, which validates the importance of the adaptive regularization. Although the prediction from lo-

Table 4: Ablation study on each component of our method.

Local Relation	Global Relation	Adaptive Ensemble	Office-Home	Office-31	VisDA	DomainNet
			71.8	88.6	82.9	67.1
✓			73.6	89.7	87.8	68.0
	✓		73.4	89.2	87.0	68.2
✓	✓		74.7	90.4	89.0	68.9
✓	✓	✓	75.2	90.6	89.8	69.9

Table 5: Analysis for the impact of data division.

Method	VisDA	DomainNet
SHOT (baseline)	82.9	67.1
Ours w/o division	87.3 (-2.5)	68.8 (-1.1)
Ours w/o class-balanced strategy	88.9 (-0.9)	69.4 (-0.5)
Ours	89.8	69.9

cal or global structure relation can not be totally accurate, we can mitigate the negative effect of noise by integrating the reliability as regularization for each branch and for each sample.

Impact of data division. We divide the target samples into two sets and establish sample connections with the source-like samples, thus to distinguish and better utilize the reliable information. In order to prove the necessity of such data division, we conduct experiments without division, where the relation is built with samples from the overall set instead of from source-like set. In Tab. 5, compared to baseline, the method without data division can still boost accuracy by a large margin, which suggests that the structural information helps with sample training. Furthermore, with the sample division, we can build more reliable local and global relationship by recognizing those source-like samples, thus exploiting the reliable knowledge and transferring such knowledge to unreliable samples. Besides, we also analysis the class-balanced division strategy, showing that by dividing samples with class-wise threshold, it alleviates learning imbalance and further promotes adaptation effect.

In Fig. 3, we give the curve of sample set proportion and model performance on VisDA during training. As the training goes, the model progressively adapts to the target domain data, leading to a gradually increasing number of reliable samples. On this basis, the established relation can take advantage of more domain information and be more robust, which in turn facilitates model adaptation and boosts performance. Finally, our method converges to a stable and higher classification accuracy with the proposed learning scheme.

Impact of distillation. In Tab. 6, we compare the results of methods with and without the teacher-student distillation scheme on VisDA. Compared to the conventional pseudo label based self-training, the employment of knowledge distillation helps the model to achieve better performance thanks to the supervision from teacher. Furthermore, our relation distillation method significantly

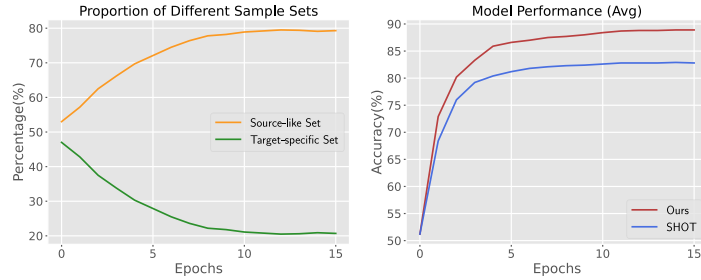


Fig. 3: Training statistics on VisDA. **Left:** The percentage of each sample set during training. **Right:** The model performance (Avg. accuracy) comparison between our method and SHOT [18].

Table 6: Effectiveness evaluation of the distillation scheme.

Distillation	Supervision Type	Accuracy(%)
$\times$	hard pseudo label	82.6
	soft pseudo label	84.5
$\checkmark$	logit distillation	85.2
	relation distillation	89.8

outperforms the traditional knowledge distillation that based on logit consistency by 4.6%, demonstrating the benefits brought by considering data structure and sample relationships.

4.4 Analysis

Superiority of the relation supervision. To validate our core insight, *i.e.*, employing inherent similarity and relationship across samples instead of conventional pseudo labels to help model training, we evaluate the prediction accuracy directly from classifier output and from the proposed hierarchical relations. Specifically, we report the results on source-like set, target-specific set and the overall dataset, as illustrated in Fig. 4. Firstly, it is observed that the prediction from local and global structural relation exhibit higher accuracy than conventional pseudo label, suggesting that it is appropriate to utilize hierarchical relation as supervision to train student model. Besides, the accuracy on source-like set is much higher than that on target-specific set. On account of these trustworthy predictions, it is reasonable to exploit source-like set and propagate these knowledge to facilitate the learning of other samples by relation distillation.

Hyper-parameter sensitivity. In this part, we analyze the model sensitivity of hyper-parameters, *i.e.*, the number of nearest neighbors K , the EMA update coefficient λ , and the temperature ρ in relation distillation. We vary these hyper-parameters, conduct thorough experiments on VisDA, and report the sensitivity in Fig. 5. First, the method achieves the best performance when we select

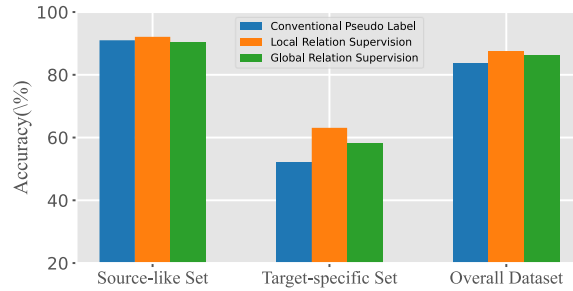


Fig. 4: The accuracy of different types of supervision, which is evaluated on VisDA.

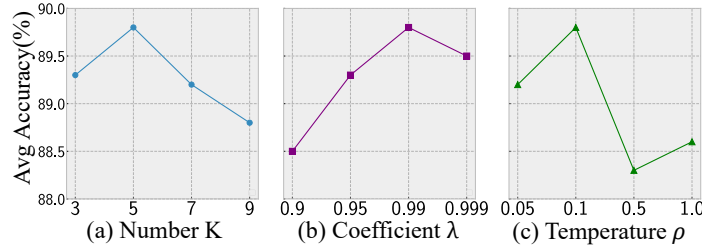


Fig. 5: Hyper-parameter sensitivity on VisDA.

the top-5 nearest neighbors for local clustering relation, while the larger K leads to performance degradation. The reason may be that including more neighbors is more likely to introduce noisy predictions and thus affect the model performance. Second, for the coefficient λ , we find it is stable when selected from $\{0.95, 0.99, 0.999\}$, which all achieve the accuracy above 89.0%, and we finally set λ as 0.99. Third, as for the temperature ρ that is used to calibrate the similarity distribution for distillation, we empirically set it as 0.1.

5 Conclusion

In this paper, we present a novel Hierarchical Relation Distillation framework, specifically designed for the challenging task of source free domain adaptation. By leveraging relation based supervision from different properties, HRD overcomes the limitations of traditional self-training approaches that neglect data structure information and inter-sample relationships, ensuring a seamless transfer of structural knowledge from teacher to student. Through the effective transfer of both local clustering and global semantic relations, the student model gains more sufficient and reliable structural knowledge from the teacher, leading to enhanced adaptability and generalization in the target domain. Extensive experiments validates the effectiveness of the proposed method, demonstrating the superior performance of HRD.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) (No. 62371009 and No. 61971008).

References

1. Bucci, S., D’Innocente, A., Liao, Y., Carlucci, F.M., Caputo, B., Tommasi, T.: Self-supervised learning across domains. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(9), 5516–5528 (2022)
2. Chen, D., Wang, D., Darrell, T., Ebrahimi, S.: Contrastive test-time adaptation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 295–305 (2022)
3. Chen, Y., Zhu, X., Li, Y., Li, Y., Wei, Y., Fang, H.: Contrast and clustering: Learning neighborhood pair representation for source-free domain adaptation. *arXiv preprint arXiv:2301.13428* (2023)
4. Fang, Y., Yap, P.T., Lin, W., Zhu, H., Liu, M.: Source-free unsupervised domain adaptation: A survey. *arXiv preprint arXiv:2301.00265* (2022)
5. Farahani, A., Voghoei, S., Rasheed, K., Arabnia, H.R.: A brief review of domain adaptation. *arXiv preprint arXiv:2010.03978* (2020)
6. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. *International Journal of Computer Vision* **129**, 1789–1819 (2021)
7. Guillory, D., Shankar, V., Ebrahimi, S., Darrell, T., Schmidt, L.: Predicting with confidence on unseen distributions. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 1134–1144 (2021)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 770–778 (2016)
9. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015)
10. Jin, Y., Wang, X., Long, M., Wang, J.: Minimum class confusion for versatile domain adaptation. In: *Proceedings of the European Conference on Computer Vision*. pp. 464–480 (2020)
11. Jing, M., Zhen, X., Li, J., Snoek, C.: Variational model perturbation for source-free domain adaptation. *Advances in Neural Information Processing Systems* **35**, 17173–17187 (2022)
12. Karim, N., Mithun, N.C., Rajvanshi, A., Chiu, H.p., Samarasekera, S., Rahnavard, N.: C-sfda: A curriculum learning aided self-training framework for efficient source free domain adaptation. *arXiv preprint arXiv:2303.17132* (2023)
13. Kundu, J.N., Bhambri, S., Kulkarni, A., Sarkar, H., Jampani, V., Babu, R.V.: Concurrent subsidiary supervision for unsupervised source-free domain adaptation. In: *European Conference on Computer Vision*. pp. 177–194 (2022)
14. Lee, J., Jung, D., Yim, J., Yoon, S.: Confidence score for source-free unsupervised domain adaptation. In: *Proceedings of the International Conference on Machine Learning*. pp. 12365–12377 (2022)
15. Li, R., Jiao, Q., Cao, W., Wong, H.S., Wu, S.: Model adaptation: Unsupervised domain adaptation without source data. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 9641–9650 (2020)

16. Li, S., Xie, M., Lv, F., Liu, C.H., Liang, J., Qin, C., Li, W.: Semantic concentration for domain adaptation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9102–9111 (2021)
17. Liang, J., He, R., Tan, T.: A comprehensive survey on test-time adaptation under distribution shifts. arXiv preprint arXiv:2303.15361 (2023)
18. Liang, J., Hu, D., Feng, J.: Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In: Proceedings of the International Conference on Machine Learning. pp. 6028–6039 (2020)
19. Liang, J., Hu, D., Wang, Y., He, R., Feng, J.: Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(11), 8602–8617 (2021)
20. Litrico, M., Del Bue, A., Morerio, P.: Guiding pseudo-labels with uncertainty estimation for source-free unsupervised domain adaptation. arXiv preprint arXiv:2303.03770 (2023)
21. Long, M., Cao, Z., Wang, J., Jordan, M.I.: Conditional adversarial domain adaptation. *Advances in Neural Information Processing Systems* **31**, 1640–1650 (2018)
22. Morerio, P., Cavazza, J., Murino, V.: Minimal-entropy correlation alignment for unsupervised deep domain adaptation. In: Proceedings of the International Conference on Learning Representations. pp. 1–12 (2018)
23. Na, J., Jung, H., Chang, H.J., Hwang, W.: Fixbi: Bridging domain spaces for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1094–1103 (2021)
24. Pei, J., Jiang, Z., Men, A., Chen, L., Liu, Y., Chen, Q.: Uncertainty-induced transferability representation for source-free unsupervised domain adaptation. *IEEE Transactions on Image Processing* **32**, 2033–2048 (2023)
25. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1406–1415 (2019)
26. Peng, X., Usman, B., Kaushik, N., Hoffman, J., Wang, D., Saenko, K.: Visda: The visual domain adaptation challenge. arXiv preprint arXiv:1710.06924 (2017)
27. Qu, S., Chen, G., Zhang, J., Li, Z., He, W., Tao, D.: Bmd: A general class-balanced multicentric dynamic prototype strategy for source-free domain adaptation. In: European Conference on Computer Vision. pp. 165–182 (2022)
28. Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.D.: Dataset shift in machine learning. Mit Press (2008)
29. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: Proceedings of the European Conference on Computer Vision. pp. 213–226 (2010)
30. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3723–3732 (2018)
31. Sun, B., Feng, J., Saenko, K.: Correlation alignment for unsupervised domain adaptation. *Domain Adaptation in Computer Vision Applications* pp. 153–171 (2017)
32. Tang, H., Chen, K., Jia, K.: Unsupervised domain adaptation via structurally regularized deep clustering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8725–8735 (2020)
33. Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S.: Deep hashing network for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5018–5027 (2017)

34. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: Proceedings of the International Conference on Learning Representations (2021)
35. Wu, C., Cao, G., Li, Y., Xi, X., Cao, W., Wang, H.: When source-free domain adaptation meets label propagation. arXiv preprint arXiv:2301.08413 (2023)
36. Xia, H., Zhao, H., Ding, Z.: Adaptive adversarial network for source-free domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9010–9019 (2021)
37. Yang, S., Wang, Y., Van De Weijer, J., Herranz, L., Jui, S.: Generalized source-free domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8978–8987 (2021)
38. Yang, S., Wang, Y., Wang, K., Jui, S., Weijer, J.v.d.: Attracting and dispersing: A simple approach for source-free domain adaptation. *Advances in Neural Information Processing Systems* **35**, 5802–5815 (2022)
39. Yang, S., Yaxing, W., van de Weijer, J., Herranz, L., Jui, S.: Exploiting the intrinsic neighborhood structure for source-free domain adaptation. *Advances in Neural Information Processing Systems* **34**, 29393–29405 (2021)
40. Yu, Z., Li, J., Du, Z., Zhu, L., Shen, H.: A comprehensive survey on source-free domain adaptation. arXiv preprint arXiv:2302.11803 (2023)
41. Zhang, L., Bao, C., Ma, K.: Self-distillation: Towards efficient and compact neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(8), 4388–4403 (2021)
42. Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., Ma, K.: Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3713–3722 (2019)
43. Zhang, Y., Wang, Z., He, W.: Class relationship embedded learning for source-free unsupervised domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7619–7629 (2023)
44. Zhang, Y., Liu, T., Long, M., Jordan, M.: Bridging theory and algorithm for domain adaptation. In: Proceedings of the International Conference on Machine Learning. pp. 7404–7413 (2019)
45. Zhang, Z., Chen, W., Cheng, H., Li, Z., Li, S., Lin, L., Li, G.: Divide and contrast: Source-free domain adaptation via adaptive contrastive learning. *Advances in Neural Information Processing Systems* **35**, 5137–5149 (2022)
46. Zhao, B., Cui, Q., Song, R., Qiu, Y., Liang, J.: Decoupled knowledge distillation. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 11953–11962 (2022)