Misalignment-Free Relation Aggregation for Multi-Source-Free Domain Adaptation

Hao-Wei Yeh[1], Qier Meng[1], Tatsuya Harada[1,2]
[1]The University of Tokyo [2]RIKEN
yeh@mi.t.u-tokyo.ac.jp, qmeng0517@gmail.com, harada@mi.t.u-tokyo.ac.jp

Abstract

In multi-source-free domain adaptation (MSFDA), it is important to effectively fuse latent features from multiple source models to improve adaptation performance on target domain. Existing works weightedly sum source-model features for fusion, which cannot fully leverage the discriminativity of features due to misaligned semantics, and is not applicable to source models with non-identical feature dimensionalities. To mitigate these issues, we propose the idea of misalignment-free relation aggregation (MFRA): instead of directly summing the features, we aggregate the similarity relationships between target samples in each source-model feature space. Specifically, for each source model, we first compute the similarities between the target sample of interest and all the other target samples. The resulting similarities are then summed along the source models to produce the aggregated similarity. To leverage the aggregated similarity in adaptation, a peer-supervised contrastive learning and an adversarial training scheme are designed to transfer discriminative information among models. The method not only effectively preserves discriminativity from each source model after summation, but also is applicable to source models with non-identical feature dimensionalities. The proposed method achieves accuracies higher or comparable to existing MSFDA methods on various cross-domain object recognition tasks. Further studies are also conducted to verify the effectiveness of aggregating inter-sample relationships, as well as the applicability of proposed method under non-identical source-model feature dimensionalities.

1. Introduction

Recently, works in source-free domain adaptation (SFDA) have developed several methods to tackle the issue of domain shift [24], which depicts the degradation in performance when applying Deep Learning models trained on one environment (source domain) to a different environment (target domain). Compared to traditional domain adaptation (DA) methods, which usually require both labeled source data and unlabeled target data during adaptation, SFDA methods only require the model trained on the labeled source data, which is called ”source model”, and the unlabeled target data during adaptation. This makes SFDA advantageous when applying to some real-world applications where labeled source data are inaccessible during adaptation due to privacy or storage issues.

While most of the SFDA works focus on adapting with single source model [20, 19, 22, 35, 32, 36, 33, 8], in some real-world applications, multiple source models, each was trained on a different source domain, can be available for adaptation. Trained on data from different source domains, each source model consists of different strength in recognizing target data. When collaborating these strength from multiple source models, it is usually able to achieve better performance than using only one of them. The variant of the setting is called multi-source-free domain adaptation (MSFDA).

To achieve collaboration among the source models, recent works in MSFDA combine the output logits or category probabilities of the source models to construct a late-fusion ensemble model [1, 9]. The ensemble model is then regarded as a single model and is trained with Information Maximization and Self-supervised Pseudo-Labeling [20, 22], which shows promising performance in single-source SFDA. To find reliable pseudo-labels for the target data, fused latent features are usually computed by combining features from multiple source models. Pseudo-labels are then inferred from the fused features via k-means-like method [1] or nearby confident samples [9].

Regarding fusing the features from multiple source models, simple weighted sum of the features is usually employed [1, 9]. However, the method raises two issues: First, it ignores the fact that the same entry of different source-model features can represent different semantic concepts. As the example in the upper row of Figure 1, for recognizing a cat, the i-th entry of the feature from the first source model may represent the concept of “eyes”, while the same

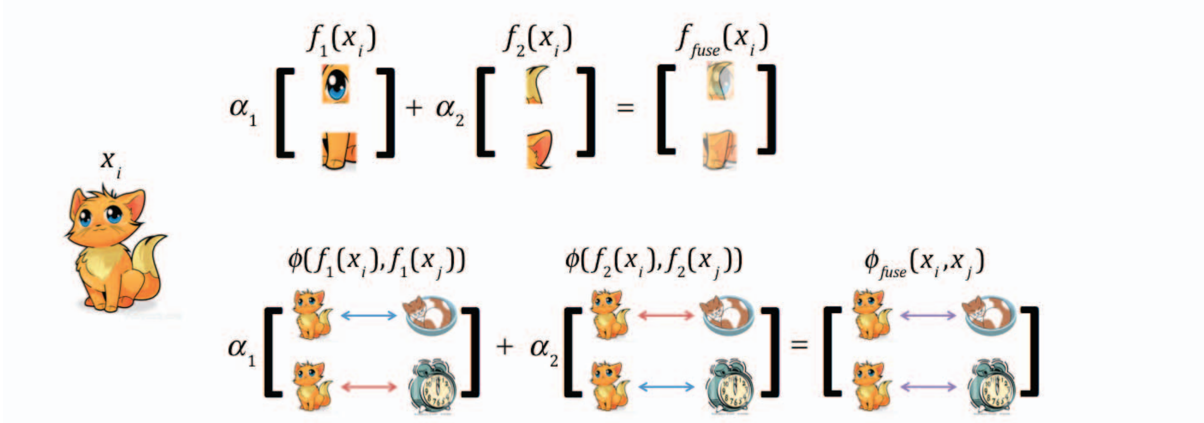


Figure 1. Comparing different fusion methods. **Upper Row** : Existing works weightedly sum $f_k(x_i)$'s, the features of target sample x_i extracted by k-th source model. Since the meanings of each entry (for example, extracting a specific part of the cat.) are not aligned, the fused feature $f_{fuse}(x_i)$ is not guaranteed to preserve original meanings after summation, which can hurt the feature discriminativity. **Lower Row** : Instead of summing features, we propose to sum the "inter-sample" relationships computed within each source model (for example, the cosine similarity $\phi(,)$ of the feature pairs). Since the meanings of each entries are aligned across the source models (for example, the j-th entry always indicates the similarity between x_i and x_j.), feature discriminativity can be better preserved after summation.

entry from the second source model may represent the concept of "tails". As we will show in the experiments, without the awareness of this misalignment, summing features from the source models could prevent us from fully leveraging the discriminativity within each of them, which we assume is due to possible destruction of the originally meaningful feature entries. Second, the method is only applicable when the dimensionalities of all features to be summed are the same, which limits its application to source models with diverse network architectures, which may then consist of non-identical feature dimensionalities. Another way to fuse source-model features is through concatenation of all features. This ensure all the information to be kept in the fused features. However, the dimensionality of the fused features will increase linearly with the number of the source models being used, which results in excessive overhead when computing with the fused features.

To mitigate the issues above, we gain our inspiration from recent advances of contrastive learning [15, 5, 3, 4, 6, 14, 2], where the training loss is usually computed with the similarity measure between features in the same feature space. Inspired by this concept, we proposed the idea of misalignment-free relation aggregation (MFRA) : instead of directly summing or concatenating source-model features, we aggregate the inter-sample relationships computed in the feature space of each source model. Specifically, within each source model, we first compute the cosine similarities of features between the target sample of interest and all the other target samples. The resulting similarities are then summed along the source models to produce the aggregated similarity. During adaptation, we train the ensemble of source models with Information Maximization and Self-supervised Pseudo-Labeling as existing MSFDA methods. In addition, we design a peer-supervised contrastive learning scheme to regularize features of each source model, with the supervision from the aggregated similarity, i.e., the supervision with the help of their "peer" models. Furthermore, to encode categorical information into features, we design an adversarial training scheme between the class predictions of the ensemble and the predictions inferred with the aggregated similarity. As the example in the lower row of Figure 1, since the similarities are computed within each source model and each the entry contains the same meaning across different source models, summing these similarities can better preserve discriminativity from each source model. Moreover, since cosine similarity produces scalar measure regardless of the dimensionality of the feature pairs, the method is also applicable to source models with non-identical feature dimensionalities.

We now summarize our contributions in this work:

- We propose the idea of misalignment-free relation aggregation (MFRA), which fuses information from multiple source models without feature misalignment, and can still be applied under non-identical dimensionalities of source-model features.

- We propose an effective framework to achieve adaptation in the setting of MSFDA via MFRA. The framework successfully transfers fused information among multiple source models in both feature and output spaces for better recognition of the target data.

- We evaluate the proposed method on various cross-domain object recognition tasks. The results show higher or comparable performance comparing to existing MSFDA methods.

2. Related Works

2.1. Multi-Source-Free Domain Adaptation

Given multiple source models, each was trained on a different source domain, and the unlabeled target data, MSFDA aims to combine the strength of multiple source models to achieve adaptation to target domain. One of the simple strategies to tackle MSFDA is to treat each source model and the target dataset as an individual pair of single-source SFDA. As an example, in the works of Source Hypothesis Transfer (SHOT) [20, 22], the proposed SFDA algorithm is applied to adapt each pair, the predictions from each pair are then averaged to produce the final prediction. However, this method neglects the possible collaboration between source models, which can be further improve the performance.

Instead of treating each source model individually, Ahmed et al. [1] introduced trainable model weights, one scalar for each source model, to weightedly sum the logits of the source models as the final predictions. The ensemble model is trained with self-supervised pseudo-labeling and information maximization as proposed in SHOT [20]. On the other hand, Dong et al. [9] also finetuned the ensemble source model, and proposed confident-anchor-induced pseudo label generator, which aims to infer reliable pseudo-labels by searching the nearest confident sample. However, regarding fusing the latent features from multiple source models, simple weighted sum of the features is employed in both works. Such method ignores the fact that the same entry of different source-model features can represent different semantic concepts, and can only apply when the features to be summed are identical in dimensionality. On the contrary, our work propose to aggregate the inter-sample relationships computed within the feature spaces of each source model, which not only keeps the same entry of the summing relationships aligned in semantic concepts, but also can be applied when the source models come with non-identical feature dimensionalities.

2.2. Contrastive Learning

In recent advances of self-supervised learning, contrastive learning achieves outstanding performance in extracting discrminative features without ground-truth labels from human labor. [15, 5, 3, 4, 6, 14, 2]. Generally speaking, contrastive learning extracts discrminative features by pulling features of the sample and the similar counterpart (positive samples) together, while pushing away the dissimilar ones (negative samples). To achieve this, a training loss is usually designed with the similarity measure between features in the same feature space. Our methods of MFRA is inspired by this concept, and we propose to aggregate the inter-sample relationships from multiple source-model feature space to fuse the discrminativity without misalignment.

In addition, finding reliable positive sample is one of the important keys to learn discrminative features in contrastive learning. In most of the self-supervised contrastive learning works, positive sample is defined as the sample itself but with a different data augmentation. Nearest-neighbor contrastive learning (NNCL) [10] extends the idea by defining reliable positive sample as the nearest neighbor of the sample of interest. On the other hand, supervised contrastive learning [17] uses ground-truth labels to define reliable positive sample as the samples having the same ground-truth label as the sample of interest. In comparison, our method conducts contrastive learning to regularize the features of each source model, with the positive sample selected by the aggregated similarity. In other words, the proposed method enables each model to leverage supervision from its "peer" models to help regularize its own feature space.

3. Proposed Method

In this section, we first explain the problem setting of MSFDA considered in this paper, then describe the underlying algorithm of the proposed method. The overall pipeline of the proposed method is summarized in Figure 2. Each part of the pipeline will be introduced in Section 3.2 to 3.6.

3.1. Problem Setting

In the problem setting of MSFDA, a set of source models $M_S = \{M_{S_1}, ..., M_{S_K}\}$ is given. Each source model M_{S_k} is a Deep Neural Network well-trained on the labeled data from the k-th source domain. For explanation purpose, we represent each source model M_{S_k} as a feature extractor F_{S_k} (parameterized by $\theta_{F_{S_k}}$) followed by a classifier C_{S_k} (parameterized by $\theta_{C_{S_k}}$), that is, $M_{S_k}(x) = (C_{S_k} \circ F_{S_k})(x)$. In addition to the source models, an unlabeled target dataset $D = \{x_i\}_{i=1}^{n}$ of size n is given. x_i represents the i-th target sample in D. The goal is to successfully infer correct class labels of target data in D by adapting models in M_S to the target domain. In this work, we consider the Close-Set DA, where the predicted categories are identical between source and target domains.

3.2. Inferring Source-model Weights

To tackle the setting of MSFDA, we start off by inferring the model weights, which indicates the confidence to each source models. With the observation that the class predictions with smaller entropy tends to be more accurate [18], we use the entropy of the class predictions to compute the model weights:

$$\alpha_k = \frac{\exp(\frac{-1}{n}\sum_i H(p_{S_{ik}})/\gamma)}{\sum_k \exp(\frac{-1}{n}\sum_i H(p_{S_{ik}})/\gamma)} \qquad (1)$$

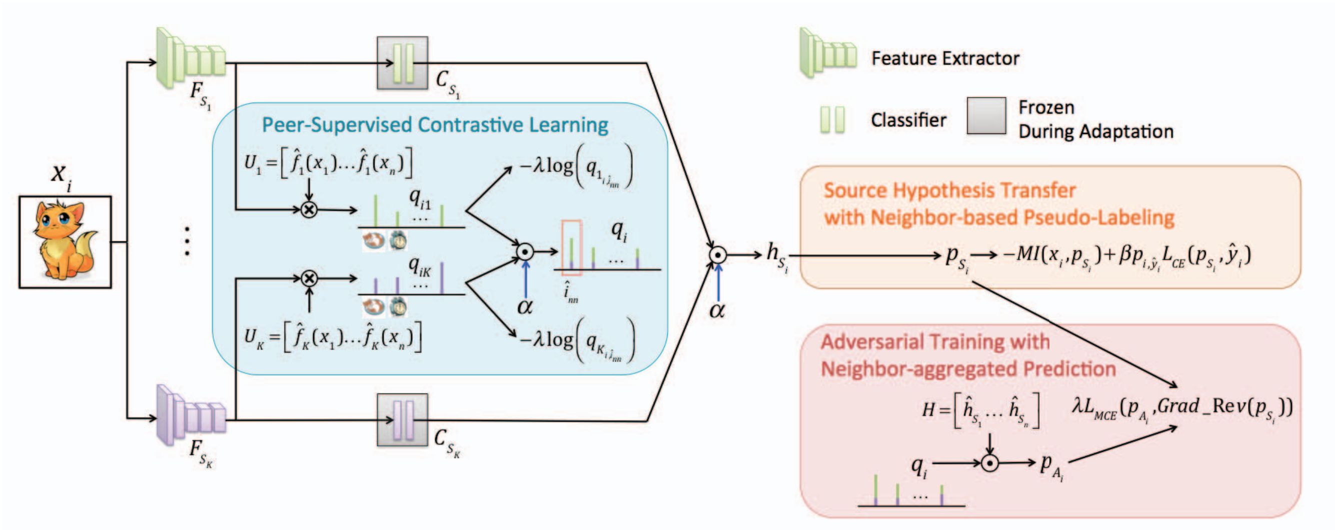

X_i
F_{S_1}
C_{S_1}
Peer-Supervised Contrastive Learning
$U_1=[\hat{f}_1(x_1)\ldots\hat{f}_1(x_n)]$
q_{i1}
$-\lambda\log\left(q_{1_{i,\hat{i}_{nn}}}\right)$
q_i
$\hat{i}_{nn}$
α
q_{iK}
$U_K=[\hat{f}_K(x_1)\ldots\hat{f}_K(x_n)]$
$-\lambda\log\left(q_{K_{i,\hat{i}_{nn}}}\right)$
F_{S_K}
C_{S_K}
h_{S_i}
α
Feature Extractor
Classifier
Frozen During Adaptation
Source Hypothesis Transfer with Neighbor-based Pseudo-Labeling
$p_{S_i} \rightarrow -MI(x_i,p_{S_i})+\beta p_{i,\hat{y}_i}L_{CE}(p_{S_i},\hat{y}_i)$
Adversarial Training with Neighbor-aggregated Prediction
$H=[\hat{h}_{S_1}\ldots\hat{h}_{S_n}]$
$\lambda L_{MCE}(p_{A_i},Grad\_Rev(p_{S_i}))$
q_i
p_{A_i}

Figure 2. The overall pipeline of the proposed method. Feature extractors F_S and classifiers C_S are initalized with parameters from the source models, each was trained with different source data. Given a target data x_i, the logits output by each source classifiers are weightedly summed with model weights α to produce the final prediction p_{S_i}. We adapt the ensemble model via Source Hypothesis Transfer [20], with the pseudo-labeling method improved by a neighbor-based strategy (orange block). Additionally, inter-sample similarities q_{k_i} are computed within each source model. The similarities are then weighted summed into aggregated similarity q_i and used to estimate the neighbor-aggregated prediction p_{A_i}. To transfer discriminativity and class semantics between source models, peer-supervised contrastive learning (blue block) is applied to each q_{k_i} and adversarial training (red block) is applied between p_{S_i} and p_{A_i}.

, where $H(p)$ is the entropy of the prediction p. γ is the hyper-parameter for scaling. If not explicitly mentioned, we set $\gamma = 0.5$ in the experiments.

For computing the final prediction p_{S_i} of the ensemble model, we follow [1] by summing logits from each source model weighted by the model weights :

$$h_{S_i} = \sum_k \alpha_k * h_{S_{k_i}} \tag{2}$$

, where $h_{S_{k_i}}$ is the output logit of source model M_{S_k} for sample x_i. The final prediction is then computed as $p_{S_i} = \sigma(h_{S_i})$, where $\sigma(.)$ is the softmax function.

3.3. Source Hypothesis Transfer with Neighbor-based Pseudo-Labeling

Next, we apply Source Hypothesis Transfer (SHOT) [20] to adapt the ensemble model as in the existing works [1, 9]: we freeze the source classifiers C_{S_k}'s and apply Information Maximization and Self-supervised Pseudo-Labeling [20] to adapt the feature extractors F_{S_k}'s to the target data. The loss function is summarized as follows :

$$L_{SHOT} = -MI(x_i, p_{S_i}) + \beta * \hat{p}_{i,\hat{y}_i} * L_{CE}(p_{S_i}, \hat{y}_i) \tag{3}$$

$MI(x,p)$ measures the mutual information between data x and class prediction p, and is computed by $\mathbb{E}_{x_i \in D}[H(p_{S_i})] - H(\mathbb{E}_{x_i \in D}[p_{S_i}])$. $L_{CE}(p,y)$ computes the cross-entropy loss between class prediction p and the corresponding label y. Here we modify the loss term to further consider $\hat{p}_{i,\hat{y}_i}$, the probability of class $\hat{y}_i$. Following the setting in [1], we set the hyper-parameter β=0.3.

Though existing methods usually infer the pseudo labels $\hat{y}_i$ via centroid-based methods [1, 9], in order to better capture the local feature structure of each source model, we modify the pseudo-labeling method to a neighbor-based strategy, which is an extension of the method designed for traditional DA [21] to MSFDA. Specifically, let N_{ik} be the set of m-nearest-neighbors of x_i under a distance measure $Dist_f$ in the feature space constructed by F_{S_k}, the class prediction $\hat{p}_i$ for x_i can be computed by :

$$\hat{p}_i = \sum_k \frac{\alpha_k}{|N_{ik}|} \sum_{x_i \in N_{ik}} p_{S_{ik}} \tag{4}$$

The pseudo-label $\hat{y}_i$ is then inferred as the most probable class of $\hat{p}_i$, i.e., $\hat{y}_i = \text{argmax}_c \, \hat{p}_i$. We use cosine distance as $Dist_f$, and set $m = 5$ as suggested in [21].

3.4. Peer-Supervised Contrastive Learning

In addition to exploiting training signals from the class predictions, we also would like to regularize the training by leveraging discriminativity that resides in the latent features of multiple source models. As described in Section 1, summing source-model features could be affected by the misalignment of semantic concepts, thus the discriminativity cannot be fully leveraged.

To better preserve and leverage discriminativity from multiple source models, instead of summing source-model features directly, we propose to aggregate the inter-sample relationships of features. Since the relationships can be computed within each source-model feature space, and the meanings of each entry of the relationships is aligned across the source models, the discriminativity can be better preserved and leveraged. Specifically, for each source model M_{S_k}, we construct a memory bank $U_k = [\hat{f}_k(x_1), ..., \hat{f}_k(x_n)]$, where $\hat{f}_k(x_i)$ represents the "cached" l2-normalized feature of target sample x_i extracted from F_{S_k} in previous iterations. We then compute the similarities q_{k_i} between the l2-normalized feature $f_k(x_i)$ and the features in memory bank U_k as follows:

$$q_{k_{i,j}} = \frac{\exp(\hat{f}_k(x_j)^T f_k(x_i)/\tau)}{\sum_{j \neq i} \exp(\hat{f}_k(x_j)^T f_k(x_i)/\tau)} \tag{5}$$

where τ is the temperature parameter and we set $\tau = 0.07$ as suggested in previous work [31]. Next, we aggregate q_{k_i}'s from each source model weighted by model weights α.

$$q_{i,j} = \sum_k \alpha_k * q_{k_{i,j}} \tag{6}$$

To transfer discriminative information among models, we use aggregated similarity q_i to infer the nearest neighbor of sample x_i, denoted as $x_{\hat{i}_{nn}}$. We then conduct the peer-supervised contrastive learning on each aggregating similarity q_{k_i}, with $x_{\hat{i}_{nn}}$ being selected as positive via aggregated similarity q_i, i.e., the help from the "peer" models, and the other samples as negatives :

$$L_{PSC} = \sum_k -\log(q_{k_{i,\hat{i}_{nn}}}) \tag{7}$$

, where $\hat{i}_{nn} = \text{argmax}_j(q_{i,j})$.

3.5. Adversarial Training with Neighbor-Aggregated Prediction

In order to leverage the learned features for classification, we also would like to encode the categorical information into the adapted features. To achieve this, we first compute a neighbor-aggregated prediction p_{A_i} based on the aggregated similarity q_i :

$$p_{A_i} = \sigma(\sum_{j=1}^{n}(H_j * q_{i,j})) \tag{8}$$

where $\sigma(.)$ is the softmax function. H is the memory bank that caches the fused logits (h_{S_i} in Equation 2) computed in previous iterations. We then conduct adversarial training between the final prediction p_{S_i} and the neighbor-aggregated prediction p_{A_i}:

$$L_{ADV} = \mathbb{E}_{x_i \in D}[L_{MCE}(p_{S_i}, Grad\_rev(p_{A_i}))] \tag{9}$$

Here $Grad\_rev(.)$ represents the gradient reversal operation. For the purpose of mitigating the effect of gradient vanishing and exploding during adversarial training [12], we use the modified cross-entropy $L_{MCE}(p_i, q_i) = \sum_c q_{i,c} \log(1 - p_{i,c})$ for the adversarial loss L_{ADV}. Intuitively, p_{S_i} maximizes L_{ADV} to corrupt the class semantics. On the other hand, p_{A_i} minimizes L_{ADV} to try to recover the class semantics by leveraging information from neighboring samples but not themselves, thus making the target samples cluster more.

3.6. Overall Objective

Finally, we summarize the overall training objective for the entire pipeline by combining the losses introduced above :

$$\hat{\theta}_{F_S} = \underset{\theta_{F_S}}{\text{argmin}}\, L_{SHOT} + \lambda(L_{PSC} + L_{ADV}) \tag{10}$$

, where θ_{F_S} are the trainable parameters for the set of feature extractors $\{F_{S_1}, ..., F_{S_K}\}$. λ is the balance hyper-parameter and we set $\lambda = 0.1$ for all the experiments.

4. Experiment Results

4.1. Setups

To evaluate the proposed method, we conduct experiments on 4 benchmark datasets for cross-domain object recognition tasks : **1) Office31** [25]: A small-scale benchmark dataset with 31 object classes from 3 domains : images from amazon.com (A), images taken in office environment with webcams (W) and DSLR cameras (D). **2) Office-Caltech** [11]: A small-scale benchmark dataset which is built by extracting 10 overlapping object classes of two datasets : Office31 [25] and Caltech256 [13]. The dataset consists of images from 4 domains : the 3 domains as in Office31 (A, W, D), and real-world images from the Caltech256 dataset (C). **3) Office-Home** [29]: A medium-scale benchmark dataset with images of 65 object classes from 4 domains : artistic images (A), clipart images (C), product images (P), and real-world images (R). **4) DomainNet** dataset [23]: A large-scale benchmark dataset with 345 object classes from 6 domains : clipart images (Cl), infograph images (In), painting images (Pa), quickdraw images (Qu), real-world images (Re), and sketch images (Sk). To ease the effect of noisy labels from some of the classes, we select the subset of 126 classes suggested in [26] for the 6 domains. For each dataset, each of the domain takes turns to be the target domain, while the rest of the domains in the dataset being the source domains.[1]

[1] Due to limited page length, we present the results on DomainNet dataset in the supplementary material.

Table 1. Accuracies (%) of object recognition on Office31 dataset.

Method	Source-free	r→A	r→D	r→W	Average
LtC-MSDA [30]	✗	68.6	99.4	97.7	88.6
M^3SDA-β [23]	✗	69.4	99.6	99.3	89.5
SImpAl [28]	✗	70.6	99.2	97.4	89.0
Source Ensemble	✓	65.9	97.3	95.5	86.2
SHOT [20]	✓	75.0	97.8	94.9	89.3
DECISION [1]	✓	75.4	99.6	98.4	91.1
CAiDA [9]	✓	75.8	**99.8**	**98.9**	**91.6**
Ours	✓	**76.4**	99.7	98.3	91.5

Table 2. Accuracies (%) of object recognition on Office-Caltech dataset

Method	Source-free	r→A	r→C	r→D	r→W	Average
MCD [27]	✗	92.1	91.5	99.1	99.5	95.6
M^3SDA-β [23]	✗	94.5	92.2	99.2	99.5	96.4
CMSS [34]	✗	96.0	93.7	99.3	99.6	97.2
Source Ensemble	✓	95.6	93.6	98.7	98.2	96.5
SHOT++ [22]	✓	96.2	96.5	99.4	**100.0**	98.0
DECISION [1]	✓	95.9	95.9	**100.0**	99.6	98.0
CAiDA [9]	✓	**96.8**	**97.1**	**100.0**	99.8	**98.4**
Ours	✓	95.8	96.3	**100.0**	99.5	97.9

Table 3. Accuracies (%) of object recognition on Office-Home dataset

Method	Source-free	r→A	r→C	r→P	r→R	Average
M^3SDA-β [23]	✗	67.2	58.6	79.1	81.2	71.5
MCD [27]	✗	69.8	59.8	80.9	82.7	73.3
SImpAl [28]	✗	72.1	62.0	80.3	81.8	74.1
Source Ensemble	✓	68.8	52.7	78.2	80.9	70.2
SHOT [20]	✓	73.0	60.4	83.9	83.3	75.2
SHOT++ [22]	✓	73.1	61.3	84.3	84.0	75.7
DECISION [1]	✓	74.5	59.4	84.4	83.6	75.5
CAiDA [9]	✓	75.2	60.5	84.7	84.2	76.2
Ours	✓	**75.7**	**63.0**	**85.7**	**85.4**	**77.4**

4.2. Implementation Details

Regarding the backbone network of the source models, ImageNet-pre-trained [7] ResNet50 [16] is used in the experiments with Office31, Office-Home, and DomainNet, and ImageNet-pre-trained [7] ResNet101 [16] is used in the experiments with Office-Caltech. We follow the network architectures in SHOT [20] and replace the original classifier in the backbone with a bottleneck layer and a classifier for the recognition tasks on each dataset. To produce source model for each source domain, the entire network is trained with the commonly used cross-entropy loss on the labeled dataset of the corresponding source domain. We follow the training scheduling provided by the code of DECISION [1] [2]. For the proposed method, given the source models, we freeze the classifiers and follow the training scheduling in [1]. We adapt the entire ensemble model on target data for 15 epochs, with momentum SGD as the optimizer and batchsize as 32. For the results of the proposed method in Table 1, 2, and 3, we follow the existing works [1, 9] and report the average accuracy over 3 runs with different random seeds.

4.3. Compared Baselines

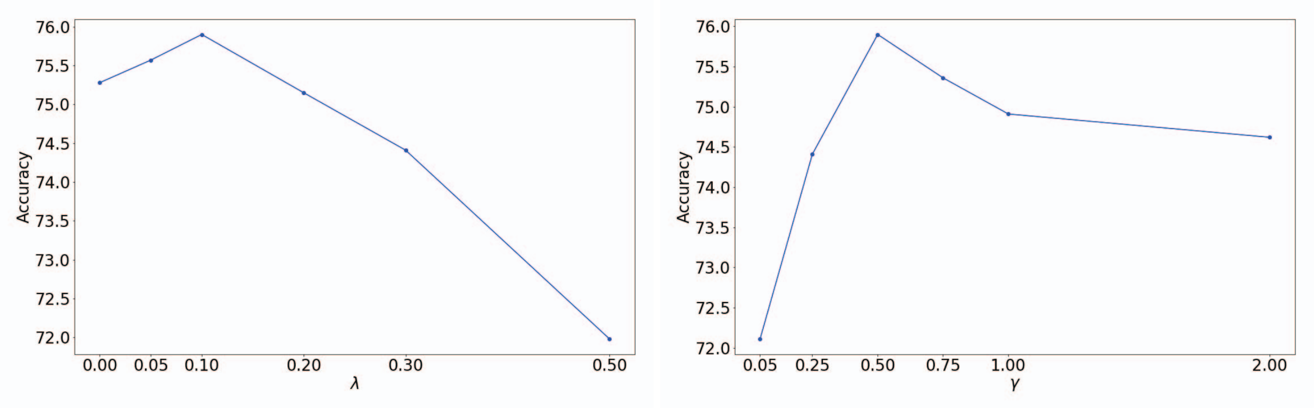
For the compared baselines, we first report the performance of **Source Ensemble** by taking the average the logits from each source model as the final prediction. For the source-free baselines, we compare the proposed method with the following existing methods : **SHOT** [20], **SHOT++** [22], **DECISION** [1] and **CAiDA** [9] [3]. In addition, we also report results from several multi-source domain adaptation algorithms that use labeled data from multiple source domains during adaptation. [30, 23, 28, 27, 34]

4.4. Results

The results of Office31, Office-Caltech, and Office-Home are summarized in Table 1, 2, and 3, respectively. We use the symbol 'r' to represent the domains excluding

[2] https://github.com/driptaRC/DECISION

[3] For SHOT and SHOT++, the adaptation is achieved by first apply the method on each source-model-target-data pair, then sum up the predictions from each pair as the final prediction



(a)

(b)

Figure 3. Sensitivity analysis of the hyper-parameter λ (a) and γ (b) in r→A experiment of OfficeHome.

the aimed target domain. As shown from the results, within the source-free group, the proposed method achieves higher or comparable accuracy compared to the existing methods. Note that SHOT and SHOT++ tackle MSFDA by considering each source model and the target dataset as an individual single-source SFDA problem. Such strategy cannot effectively collaborate multiple source model in recognizing the same target dataset, thus achieves suboptimal performance.

On the other hand, CAiDA and DECISION tackle MSFDA by constructing the ensemble of the given source models, which can better collaborate multiple source model than considering them individually. However, when fusing multiple features, both works weightedly sum features from multiple models, which can destroy the discriminativity of features and degrade the quality of the inferred pseudo-labels. Instead of directly summing the features themselves, our method aggregates the similarity between target samples that are computed within the feature space of each source model. Such method can better preserving discriminativity of source-model features, and thus achieves better adaptation performance.

In summary, compared to existing MSFDA methods, the proposed method achieves comparable and better performance. The various scales of the datasets also suggest the proposed method is applicable not only in simple but also difficult scenarios.

5. Discussions

5.1. Ablation Study

Here we provide the ablation study of the proposed method by removing each of the proposed components from the training objective. The results on the r→A setting of Office-Home are summarized in Table 4. As we can see from the table, the best result is achieved by using all the proposed components, which verifies the necessity of each of them, and shows the effectiveness of fusing features through similarity aggregation, as well as transferring feature and class semantic supervisions among source models.

Table 4. Accuracies (%) of ablation study for the proposed method

Method			Office-Home (r→A)
L_{SHOT}	L_{PSC}	L_{ADV}	
✓			75.3
✓	✓		75.8
✓		✓	75.4
✓	✓	✓	**75.9**

5.2. Sensitivity Analysis

Here we provide the sensitivity analysis on the hyper-parameter λ in the objective (Equation 10). We change the value of λ in the set $\{0, 0.05, 0.1, 0.2, 0.3, 0.5\}$ and observe the change in accuracy in the r→A experiment of Office-Home. The results are summarized in Figure 3 (a). When changing λ in small values (from 0 to 0.1), we observe improvements in accuracy, which indicates the importance of peer-supervised contrastive learning and adversarial training. The accuracy reaches its best when $\lambda = 0.1$ and drops for larger values. This maybe because that large λ induces drastic changes of loss in adversarial training, making the training unstable and desired class semantics are fail to be transferred among source models. Therefore, small values of λ are recommended for the proposed method.

We also provide the sensitivity analysis on the hyper-parameter γ that controls the scaling of the model weights. We change the value of γ in the set $\{0.05, 0.25, 0.75, 1.0, 2.0\}$ and again observe the change in accuracy in the r→A experiment of Office-Home. The results are summarized in Figure 3 (b). When γ is too small, the model weights α become too peaky and the performance is degraded due to the source models cannot be well collaborated. On the other hand, when γ is too large, the α become too smooth thus the models with noisy predictions cannot be effectively penalized, which again degrades the performance. The accuracy reaches its best when $\gamma = 0.5$.

Table 5. Accuracies (%) of object recognition on Office-Home dataset under non-identical feature dimensionalities

Method	r→A	r→C	r→P	r→R	Average
Source Only	68.6	51.9	78.7	81.2	70.1
DECISION [1]	74.0	58.5	81.7	83.2	74.4
Ours	**75.8**	**62.5**	**85.4**	**85.4**	**77.3**

Table 6. Success retrieval rate (%) of different feature spaces on r→A setting of Office-Home dataset

Feature space	Success retrieval rate on target samples
Source model C	62.9
Source model P	65.6
Source model R	71.0
Summing features	71.3
Summing relations	**73.0**

5.3. MSFDA Under Non-identical Feature Dimensionalities

In some real-world applications, due to various concerns when each source model was trained, it is possible that the feature dimensionalities of each source model are not exactly identical. As a reliable MSFDA algorithm, it is important to be applicable even under such condition. Here we evaluate the proposed method with source models having diverse feature dimensionality. We conduct experiments on the Office-Home dataset, with the feature dimensionalities of the three source models being set to 512, 256, and 256, respectively, and compare the performance with DECISION [1]. Since weightedly summing the features is not applicable in this case, we disable the pseudo-labeling method in DECISION and use the rest of the losses to adapt the model. The results are summarized in Table 5, which shows the proposed method achieves higher accuracy compared to DECISION. This implies the advantage of the proposed method that it is applicable even when the feature dimensionalities of the given source models are not exactly identical.

5.4. Nearest-Neighbor Retrieval with Different Fused Feature Spaces

In Section 1, we claimed the advantage of summing inter-sample relationship over summing features, which motivates us to propose the method of MFRA. To further investigate the claim, we evaluate the resulting feature spaces with the experiment of nearest-neighbor retrieval : Given a query sample, we find the nearest neighboring sample in the evaluated feature space. The retrieval is called a "success" if the returned sample belongs to the same class as the query sample. We use the r→A setting of Office-Home for the experiment. Given the target features extracted from the source models, we compute the fused feature space with two methods : 1) "Summing features" : weighted-summing features 2) "Summing relations": weighted-summing the inter-sample distance [4]. We evaluate the success retrieval rate of each individual and fused feature spaces on the target dataset. The results are summarized in Table 6. As shown in the table, both fusion method improves the retrieval rate, where "Summing relations" achieves better performance than "Summing features". This provides the reason why the proposed MFRA can achieve higher and comparable performance comparing to the existing methods.

6. Conclusion

In this paper, we proposed the idea of misalignment-free relation aggregation (MFRA) for multi-source-free domain adaptation (MSFDA). Unlike existing MSFDA methods, which directly sum the feature extracted from multiple source models, our method aggregates the similarity measures computed between features to fuse information from multiple source models. The proposed strategies not only can better preserve the discriminativity of features from multiple source models, but also can be applied when the source models come with non-identical feature dimensionalities. To perform adaptation, in addition to the information maximization and pseudo-labeling objective that was commonly used in existing MSFDA methods, based on the aggregated similarity, we design a peer-supervised contrastive learning and an adversarial training scheme to transfer discriminative information among different source models, which further improves the performance of the ensemble of source models. We evaluate the proposed method on various cross-domain object recognition tasks and achieve higher or comparable accuracy comparing to existing MSFDA methods. Further studies with nearest-neighbor retrieval are conducted to verify the advantage of aggregating inter-sample relationships. The evaluation with non-identical feature dimensionalities also suggests that the proposed method is still applicable when the feature dimensionalities of the given source models are not exactly identical.

7. Acknowledgements

We want to thank Yusuke Mukuta and Thomas Westfechtel for inspiring discussions, and our reviewers for helpful feedback. This work was partially supported by JST Moonshot R&D Grant Number JPMJPS2011, CREST Grant Number JPMJCR2015 and Basic Research Grant (Super AI) of Institute for AI and Beyond of the University of Tokyo.

[4] Here we use the cosine distance of the features

References

[1] Sk Miraj Ahmed, Dripta S Raychaudhuri, Sujoy Paul, Samet Oymak, and Amit K Roy-Chowdhury. Unsupervised multi-source domain adaptation without access to source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10103–10112, 2021. 1, 3, 4, 6, 8

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020. 2, 3

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 3

[4] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020. 2, 3

[5] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2, 3

[6] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 2, 3

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

[8] Ning Ding, Yixing Xu, Yehui Tang, Chao Xu, Yunhe Wang, and Dacheng Tao. Source-free domain adaptation via distribution estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7212–7222, June 2022. 1

[9] Jiahua Dong, Zhen Fang, Anjin Liu, Gan Sun, and Tongliang Liu. Confident anchor-induced multi-source free domain adaptation. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 1, 3, 4, 6

[10] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9588–9597, 2021. 3

[11] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2066–2073. IEEE, 2012. 5

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 5

[13] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007. 5

[14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020. 2, 3

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2, 3

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

[17] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020. 3

[18] Youngeun Kim, Donghyeon Cho, Kyeongtak Han, Priyadarshini Panda, and Sungeun Hong. Domain adaptation without source data. *IEEE Transactions on Artificial Intelligence*, 2(6):508–518, 2021. 3

[19] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9641–9650, 2020. 1

[20] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. *arXiv preprint arXiv:2002.08546*, 2020. 1, 3, 4, 6

[21] Jian Liang, Dapeng Hu, and Jiashi Feng. Domain adaptation with auxiliary target domain-oriented classifier. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16632–16642, 2021. 4

[22] Jian Liang, Dapeng Hu, Yunbo Wang, Ran He, and Jiashi Feng. Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1, 3, 6

[23] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1406–1415, 2019. 5, 6

[24] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010. 1

[25] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010. 5

[26] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation

via minimax entropy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8050–8058, 2019. 5

[27] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018. 6

[28] Naveen Venkat, Jogendra Nath Kundu, Durgesh Singh, Ambareesh Revanur, et al. Your classifier can secretly suffice multi-source domain adaptation. *Advances in Neural Information Processing Systems*, 33:4647–4659, 2020. 6

[29] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *(IEEE) Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5

[30] Hang Wang, Minghao Xu, Bingbing Ni, and Wenjun Zhang. Learning to combine: Knowledge aggregation for multi-source domain adaptation. In *European Conference on Computer Vision*, pages 727–744. Springer, 2020. 6

[31] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 5

[32] Haifeng Xia, Handong Zhao, and Zhengming Ding. Adaptive adversarial network for source-free domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9010–9019, October 2021. 1

[33] Baoyao Yang, Hao-Wei Yeh, Tatsuya Harada, and Pong C Yuen. Model-induced generalization error bound for information-theoretic representation learning in source-data-free unsupervised domain adaptation. *IEEE Transactions on Image Processing*, 31:419–432, 2021. 1

[34] Luyu Yang, Yogesh Balaji, Ser-Nam Lim, and Abhinav Shrivastava. Curriculum manager for source selection in multi-source domain adaptation. In *European Conference on Computer Vision*, pages 608–624. Springer, 2020. 6

[35] Shiqi Yang, Joost van de Weijer, Luis Herranz, Shangling Jui, et al. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. *Advances in Neural Information Processing Systems*, 34, 2021. 1

[36] Hao-Wei Yeh, Baoyao Yang, Pong C Yuen, and Tatsuya Harada. Sofa: Source-data-free feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 474–483, 2021. 1