Knowledge-Based Systems 275 (2023) 110728

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

UC-SFDA: Source-free domain adaptation via uncertainty prediction and evidence-based contrastive learning

Dong Chen [a], Hongqing Zhu [a,*], Suyi Yang [b]

[a] School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China
[b] Department of Mathematics, Natural, Mathematical & Engineering Sciences, King's College London, Strand, London WC2R 2LS, United Kingdom

ARTICLE INFO

Article history:
Received 5 October 2022
Received in revised form 9 June 2023
Accepted 12 June 2023
Available online 17 June 2023

Keywords:
Evidence theory
Contrastive learning
Source-free domain adaptation
Uncertainty prediction
High-confidence

ABSTRACT

Most unsupervised domain adaptation approaches learn domain-invariant features assuming that source and target domain data are available simultaneously. In practice, the availability of source samples is only sometimes possible. This paper establishes a novel source-free domain adaptation (SFDA) framework based on uncertainty prediction and a neighborhood-guided evidence-based contrastive learning scheme. First, we develop an evidence analyzer based on the uncertainty prediction principle of Dempster–Shafer (D–S) evidence theory, which improves the network capability for discriminating different types of samples. The transformer layer with a self-attention module is adopted to capture long-distance feature dependencies such that the proposed network has better generalization ability on multiple domains. Then, we offer a high-confidence target domain sample (HCS) acquisition strategy through evidence theory, entropy criterion, and distance information. A joint confidence enhancement scheme obtains the final HCS that generates pseudo-labels. Finally, we propose an optimization method based on evidence theory, evidence-based comparative learning, and internal neighborhood structure to ensure the separability between classes and compactness within categories. Experimental results show that the proposed framework performs superiorly on two standard datasets on multiple adaptation tasks. The code for this project is available at github.com/oolown/UC-SFDA.

1. Introduction

Convolutional neural networks (CNNs) have recently been successfully applied in various computer vision tasks. Nevertheless, their performance relies heavily on the use of labeled datasets, with the training and test data being independent and identically distributed. The distribution gap between data from different domains is an essential factor affecting the mode's generalization ability. Existing unsupervised domain adaptation (UDA) seeks to reduce the domain gap between the labeled source and unlabeled target domain by aligning their data distributions.

UDA requires access to the source domain during training. Due to personal or commercial protection, users can only access pre-trained source models without source data. Recent studies have attempted to explore source-free domain adaptation (SFDA), where only an unlabeled target domain dataset and a pre-trained source model are available. Existing SFDA studies can be divided into two categories: data generation methods and model fine-tuning approaches. The former uses a pre-trained source model to regenerate source-like images (or features) [1] and achieve cross-domain adaptation by applying standard UDA techniques. However, these methods introduce additional frameworks (e.g., generators and discriminators), which may require more computation resources. Instead of generating source-like data for UDA, some studies have attempted to fine-tune a pre-trained source model by exploiting unlabeled target data in a self-supervised training scheme, which our methods seek.

Model fine-tuning methods utilize a pre-trained source model in the self-supervised process and instruct the target domain data to cluster according to category self-referentially. These methods take less time and achieve acceptable adaptation results. However, self-referencing strategies heavily depend on the accuracy of pseudo-labels. Most previous pseudo-labeling acquisition approaches are based on the feature distances between class centroids and target samples. For example, Liang et al. [2] introduced a self-supervised scheme to obtain pseudo-labels through multiple cycles. They used weighted K-means to cluster all target domain data and found each class's centroid. Tang et al. [3] used semantic information fusion to optimize the pseudo-labels using their neighbor information. The clustering results may be inaccurate if the class centroids of the target domain are computed

* Corresponding author.
E-mail address: hqzhu@ecust.edu.cn (H. Zhu).

https://doi.org/10.1016/j.knosys.2023.110728
0950-7051/

directly using all target domain features. This is because the unadapted pre-trained source model's target domain feature learning ability is weak, leading to inaccurate pseudo-label estimation of target samples. Utilizing all these unreliable pseudo-labels to update the model and compute class centroids would lead to poor aggregation and hinder the domain adaptation process.

Although CNN-based models have been widely used in various classification tasks, they do not account for the inherent uncertainty in the data related to category probabilities, where incorrect classification under uncertainty may lead to decision risk. Dempster–Shafer's (D–S) evidence theory is an effective decision-making method for expressing uncertain and unknown information. Early uncertainty quantification approaches in the computer vision community are mainly based on Bayesian theory. Recently, algorithms incorporating evidence theory into CNNs have been reported to model uncertainty in class probabilities. These approaches are referred to as a deep evidence network where a ReLU replaces the Softmax to ascertain nonnegative output, which is taken as the evidence vector for the predicted Dirichlet distribution [4].

Source-free domain adaptation cannot access the source domain data, so the target domain feature cannot be aligned with the source domain feature trained in a supervised way for domain adaptation. Moreover, since the target domain data are unlabeled, self-supervised learning is adopted to solve the SFDA algorithm using model fine-tuning methods. As mentioned, the current mainstream SFDA algorithm based on model fine-tuning mainly assigns pseudo-labels to target samples using a pre-trained source model. Such an accurate assignment usually suffers from incorrect classification in highly uncertain or lack of uncertainty prediction. This paper believes that the D–S evidence theory may solve these problems. By calculating the uncertainty of samples, the deep evidence network can be used to classify the uncertain data and screen the samples with the correct pseudo-labels, which helps to improve the SFDA model's adaptation performance.

This paper addresses a novel self-supervised model fine-tuning method to accomplish source-free domain adaptation via uncertainty prediction and evidence-based contrastive learning (UC-SFDA). First, this paper presents an evidence analyzer based on evidence theory to achieve uncertainty prediction and measure the uncertainty of CNN predictions so that the model can classify uncertain data and screen samples with correct pseudo-labels, namely, high confidence target domain samples, which largely improves the accuracy of pseudo-labels. In addition, this study develops a novel high-confidence target domain sample (HCS) acquisition strategy. With this scheme, we assign pseudo-labels to all target domain samples. Concretely, after obtaining all the sample features of the target domain, we do not use them for clustering. Instead, we first use evidence theory, entropy criterion, and feature distance to evaluate the confidence of samples. We then screen out the HCS according to the designed joint confidence enhancement scheme. Then, the pseudo-labels using the distance between sample features and the confidence class centroid features are obtained, which allows for more reliable class adjustment during the adaptation process. Second, a transformer layer with a self-attention module is adopted to capture long-distance feature dependencies and improve the network's generalization ability on other domains. Third, the HCS update intermittent training scheme continuously updates the target domain model based on the newly obtained HCS and pseudo-labels. We introduce an evidence-based contrastive learning method based on high confidence target domain samples and corresponding pseudo-labels obtained by uncertainty prediction. It is an effective self-supervised optimization scheme that fully mines the neighborhood information in the feature space to fine-tune the model to adapt to target domain data. Specifically, we develop evidence loss and neighborhood-guided evidence-based contrast loss as additional self-supervised optimization schemes to encourage global clustering and balance cluster assignment. The former provides more detailed estimates of probability distributions and helps the network to generate more discerning features, and the latter models inter-class separability and intra-class compactness. We evaluate UC-SFDA on several benchmark datasets and prove that it is superior to many methods on various domain adaptation tasks.

The main contributions of this paper are as follows:

- We propose a new self-supervised SFDA framework incorporating an evidence analyzer and transformer layer. To our knowledge, this is the first attempt to report SFDA using the uncertainty prediction of evidence theory.
- Through sample uncertainty prediction, we develop an HCS acquisition strategy based on evidence theory, entropy criterion, and feature distance.
- Pseudo-labels are assigned based on the HCS obtained by a designed joint confidence enhancement scheme. This reduces the possibility of false clustering and helps accurately adjust the categories during the adaptation process.
- We propose a neighborhood-guided evidence-based contrastive learning scheme to ensure the separability between classes and compactness within categories.

2. Related works

2.1. Source-free domain adaptation

In SFDA-related works exploring domain adaptation through pre-trained source models, the pseudo-label based methods have achieved satisfactory results due to their simplicity of operation. For example, Du et al. [5] generated a pseudo-source domain from the target data and performed distribution alignment. Qiu et al. [6] developed a source domain pseudo-prototype and target pseudo-labels for domain alignment. Liang et al. [2] proposed a target domain-specific feature extraction framework that provides a target data representation consistent with the source data through self-supervising pseudo-labels and information maximization. In addition, some works develop SFDA by constructing a virtual domain. For example, Tian et al. [7] introduced a pre-trained source model to generate a virtual space based on the Gaussian mixture model. The data distribution of this virtual domain is similar to that of the source domain. Recently, Yang et al. [8] introduced a self-supervised knowledge distillation method and adopted transformer with target pseudo-labels to improve the performance of SFDA. Huang et al. [9] presented a contrastive learning scheme to compensate for missing source data. Yang et al. [10] performed SFDA by encouraging interacting neighbors to collaborate in their label predictions.

2.2. Evidence theory

Despite the excellent performance of deep learning methods, there are still areas for improvement in quantifying and predicting uncertainty. Evidence theory is an effective method for expressing uncertain and unknowable information for decision-making. Recently, evidence theory began to be applied to deep neural networks (DNNs) to handle uncertainties associated with category probabilities to improve classification precision. For example, Huang et al. [11] introduced a lymphoma segmentation model using evidence theory and the U-Net module. Yuan et al. [12] utilized evidence theory to measure the uncertainty of the predicted output of a DNN and proposed a method to classify uncertain data. Tong et al. [13] introduced evidence deep

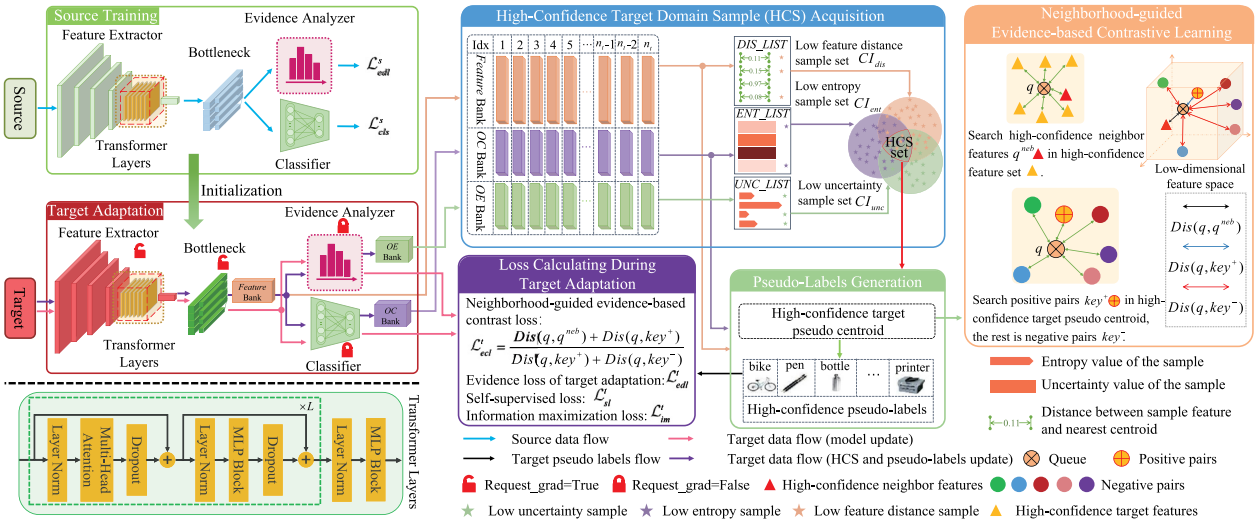


Fig. 1. The framework of the proposed UC-SFDA comprises the source training stage, target adaptation stage, HCS acquisition, and the calculation of loss functions at the target adaptation stage using evidence theory, information entropy criterion, and distance measure. Zoom in for the best view.

learning theory and CNN architecture for the set-valued classifier. The features are transformed into mass functions and fused by Dempster's rule. In the proposed UC-SFDA, an evidence analyzer is designed based on evidence theory. Compared with standard classifiers, our method uses the uncertainty of prediction results to make decisions and improves the discrimination of samples of different categories.

2.3. Contrastive learning

Contrastive learning is a form of unsupervised learning that relies on the data itself. Features are learned by attracting semantically similar samples and repelling semantically different samples. Several recent works have used contrastive learning for domain adaptation. Oord et al. [14] proposed the well-known InfoNCE, which introduces a contrastive learning framework for classification. Qiu et al. [6] performed class-wise contrastive learning to align target data with the corresponding source feature prototype. Dai et al. [15] utilized contrastive learning to maximize inter-class distance and minimize intra-class space. Huang et al. [9] used a contrastive learning strategy to compensate for the insufficiency of the source samples. In HCL [9], contrastive learning is combined with memory-based learning to solve SFDA. The contrastive learning mechanism we design combines evidence theory and exploits the intrinsic neighborhood structure of target features.

2.4. Vision transformer

Transformer [16] was first proposed primarily for machine translation in natural language processing (NLP). Recently, it has been applied to UDA due to its representative ability and inherent global attention mechanism. In [17], Sun et al. presented a self-refinement scheme for transformer-based domain adaptation and demonstrated its powerful transferable feature representation capability. Xu et al. [18] introduced a weight-sharing three-branch transformer network CDTrans for UDA. It is a bidirectional center-aware labeling scheme to generate pseudo-labels for target data. Zhang et al. [19] introduced self-attention to encourage the UDA to learn more domain transferable feature representations, mitigating the negative transfer. Yang et al. [8] used transformer to focus on object regions and further used self-supervised pseudo-labeling and knowledge distillation technologies to improve the domain generalization of the network. Unlike Yang et al. [8], the proposed UC-SFDA adopts transformer's self-supervised mechanism and evidence theory to facilitate the network's attention to objects and designs a neighborhood-guided evidence-based contrastive learning scheme for SFDA based on uncertainty prediction.

3. Methodology

To formulate the SFDA problem, this study defines the labeled source domain data as $\mathcal{D}_s = \{(x_i^s, y_i^s) \mid 1 \le i \le n_s\}$, where y_i^s is the corresponding label of sample x_i^s, and the unlabeled target domain data is denoted as $\mathcal{D}_t = \{x_j^t \mid 1 \le j \le n_t\}$. In addition, the target domain and the source domain have K numbers of identical categories. Under the SFDA setting, only the pre-trained source model ϕ_s and target domain $\mathcal{D}_t$ are available.

3.1. Overall architecture

Fig. 1 shows the entire network architecture of UC-SFDA. Compared with the classical SFDA model, the proposed framework contains two other components: transformer layer T_s which helps the feature extractor $\mathcal{F}_s$ pay more attention to objects, and an evidence analyzer E_s which describes the predicted uncertainty of a sample to improve the discrimination between different categories.

As shown in Fig. 1, The algorithm contains two stages: (i) The first stage is the source training stage. At this stage, we train a novel established network to obtain a pre-trained source model under the SFDA settings. (ii) The second stage is the target adaptation stage. At this stage, we use a HCS update intermittent training scheme to adapt the pre-trained source model to an unlabeled target domain. This training scheme alternately applies two processes, "HCS and pseudo-labels update" and "model update". Firstly, we utilize the pre-trained source model to initialize the target model and freeze the classifier C_t and evidence analyzer E_t, the entire model is not updated; we only input the target data to the target model to establish *OE* bank, *OC* bank and *Feature* bank. These three memory banks are fed into the two

modules "High Confidence Target Domain Sample (HCS) Acquisition" and "Pseudo-labels Generation" in this process to calculate HCS and corresponding pseudo-labels, respectively. Then, in the "model update" process, target samples are sent to the target model by mini-batch to produce feature output, evidence output and classification output, and then the output is fed to modules "Neighborhood-guided Evidence-based Contrastive Learning" and "Loss Calculating During Target Adaptation". With the obtained HCS set, high-confidence target pseudo centroid and high-confidence pseudo-labels during "HCS and pseudo-labels update" process, the four loss terms (Neighborhood-guided evidence-based contrast loss, Evidence loss of target adaptation, Self-supervised loss, and Information maximization loss) making up the target loss $\mathcal{L}_{tar}$ are calculated, the feature extractor $\mathcal{F}_t$ and bottleneck G_t in the target model is updated based on $\mathcal{L}_{tar}$.

3.2. Source training

At this stage, the labeled source data $\mathcal{D}_s$ are input into the network for training to obtain a pre-trained source model ϕ_s. Compared with the classical SFDA model, the proposed framework contains two other components: a transformer layer T_s which lets feature extractor $\mathcal{F}_s$ pay more attention to the object, and an evidence analyzer E_s for describing the predicted uncertainty of a sample to improve the discrimination between different categories. When the source data pass through the network, the output contains two parts: one is classifier output $O_c^s = oc_i^s = C_s(G_s(\mathcal{F}_s(x_i^s)))$, and the other is the evidence output $O_e^s = oe_i^s = E_s(G_s(\mathcal{F}_s(x_i^s)))$.

This study utilizes the cross-entropy loss with the label smoothing technique [20] to train this model, formulated as:

$$\mathcal{L}_{cls}^s = -\frac{1}{n_s}\sum_{i=1}^{n_s}\sum_{k=1}^{K} l_{i,k}^{ls}\log \delta_k(oc_i^s), \tag{1}$$

where $\delta_k(\cdot)$ represents the Softmax operation, $l_{i,k}^{ls}$ is the kth element of the smooth label $l_i^{ls} = (1-\partial)l_i + \partial/K$, ∂ is the smooth coefficient, which is set to 0.1 [20], and l_i is a one-hot encoding of label y_i^s.

Probability theory cannot describe uncertainty and can only use a particular interval value [0,1] to show the degree of support for the proposition. It could be explained as utilizing the softmax function in the cross-entropy loss, which assigns a point estimate for the category probability of each sample [21]. CNNs usually use Softmax to activate the final classifier; using exponents will amplify the probability of predicted classes. This study develops an evidence analyzer after bottleneck G_s, which leverages evidence theory to simulate the probability distribution of different classes.

Evidence theory is an extension of Bayesian theory to subjective logic. As an improvement and development of probability theory, it is an effective decision-making method to express uncertain and unknown information. Therefore, this paper innovatively introduces evidence loss based on evidence theory and adopts evidence theory in classification networks. This could solve the general problem in uncertainty classification that leads to unreliable results and improve the classification accuracy of the network.

Suppose CNN can form an opinion of classification task as a Dirichlet distribution. We need first to design an evidence analyzer and use evidence loss to train the network, and then classify a given sample i into the Dirichlet distribution $D(\mathbf{p}_i|\theta_i^s)$, where $\mathbf{p}_i = [p_{i1}, \ldots, p_{iK}]$ is the probability mass function, satisfying $\sum_{k=1}^{K} p_{i,k} = 1$, $0 \le p_{i,1}, \ldots, p_{i,K} \le 1$, θ_i^s represents the parameters of the Dirichlet distribution obtained using the evidence analyzer during the source training stage. oe_i^s feeds the ReLu activation layer $\varphi(\cdot)$ to obtain $e_i^s = \varphi(oe_i^s)$ as the evidence for the analyzer. The parameters of the Dirichlet distribution can be obtained by $\theta_i^s = e_i^s + 1$. We define the evidence loss of source training as follows:

$$\mathcal{L}_{edl}^s = \sum_{i=1}^{n_s}\sum_{k=1}^{K}\mathbb{E}[(l_{i,k} - e_{i,k}^s)^2] + \lambda_t \sum_{i=1}^{n_s} KL[D(e_i^s \mid \theta_i^s) \,||\, D(e_i^s \mid \mathbf{I})], \tag{2}$$

where t is the number of training epochs, $T_{\max}$ is the maximum training epoch, $\lambda_t = \min(1.0, t/T_{\max}) \in [0,1]$ is the annealing coefficient, and $e_{i,k}^s = \varphi_k(oe_i^s)$ is the kth element of the K-dimensional evidence output vector. $D(e_i^s \mid \mathbf{I})$ is the uniform Dirichlet distribution, where $\mathbf{I} = \langle 1, \ldots, 1\rangle$ is K-dimensional. KL represents the Kullback–Leibler (KL) divergence between two distributions. The first term of (2) achieves both objectives of minimizing: the variance of the Dirichlet distribution and the prediction error generated by the network specifically for each training sample. The second term is a regularization term that normalizes the predicted distribution by penalizing those divergences from the state that do not contribute to uncertainty prediction. During training, the evidence analyzer discovers modes in data and generates class-specific label evidence in terms of these modes to optimize the overall loss. Using evidence loss (2), the proposed module provides more detailed estimates of probability distributions than point estimates of the softmax classifier. Compared with a classical individual classifier, the proposed incorporating evidence analyzer E_s along with the classical classifier C_s can generate evidence information from the analyzer outputs. Overall, the total loss function used for the training source model is

$$\mathcal{L}_{src} = \mathcal{L}_{cls}^s + \mathcal{L}_{edl}^s. \tag{3}$$

3.3. Target adaptation

The target adaptation stage assumes that only a pre-trained source model is available and that the model is initialized as a target domain model $\phi_t = \{\mathcal{F}_t, G_t, E_t, C_t\}$. To retain the class distribution information and the estimation pattern of the pre-trained source model to quantify the belief mass and uncertainty of input samples, classifier C_t and evidence analyzer E_t remain frozen during the target adaptation phase. This study proposes an HCS update intermittent training scheme that the first acquires HCS with the help of analyzed uncertainty information to produce their pseudo-labels during HCS and pseudo-labels update and then updates the target model based on these obtained pseudo-labels in an alternating mode. Regarding the advantages of this adaptation scheme than general target adaptation methods, switching from model update back to HCS and pseudo-labels update to regenerate memory banks allows sample selection with higher confidence pseudo-label prediction. Alternating these two processes helps improve the classification precision. We introduce the alternating update between HCS and the target network model in the following. As shown in Fig. 1, Requset_grad=True and Requset_grad=False are used to represent the updated or non-updated status of network components respectively.

One epoch training for the target model contains two processes: (i) HCS and pseudo-labels update (purple data flow shown in Fig. 1). We feed the target data into the target model to establish three memory banks: *feature* bank F stores all target domain features, *OE* bank, and *OC* bank stores the corresponding evidence analyzer outputs and classifier outputs, respectively. Operated equivalently to the test phase, the network in the first process is not updated. Hence, the entire target domain network is frozen, denoted as Request_grad=False. Then, we obtained three

memory banks and used them to calculate HCS and pseudo-labels. (ii) Model update (pink data flow shown in Fig. 1): We update the target model to achieve domain adaptation. At this time, the target sample is sent to the target model by mini-batch, and $\mathcal{L}_{tar}$ in (14) is calculated to update the network. In this process, only feature extractor $\mathcal{F}_t$ and bottleneck network G_t are updated, represented by an unlocked symbol in Fig. 1. The other two partitions C_t and E_t are not updated, which is represented by a locked symbol in the figure. Therefore, feature extractor $\mathcal{F}_t$ and bottleneck G_t switch between updating (the process of model update) and remaining static (the process of HCS and pseudo-labels update). We set the whole epoch for target adaptation as T_{max}. In terms of the advantage of this training scheme, switching model update back to HCS and pseudo-labels update to regenerate memory banks allows sample selection with higher confidence pseudo-label prediction. Alternating these two processes helps improve the classification precision. To show the architecture of the HCS, we first express F, OE, and OC as

$$\begin{aligned} F =&[\bar{f}_1^t, \bar{f}_2^t, \ldots, \bar{f}_{n_t}^t], \quad OE = [o\bar{e}_1^t, o\bar{e}_2^t, \ldots, o\bar{e}_{n_t}^t], \\ &OC = [o\bar{c}_1^t, o\bar{c}_2^t, \ldots, o\bar{c}_{n_t}^t], \end{aligned} \tag{4}$$

where $\bar{f}_j^t = G_t(\mathcal{F}_t(x_j^t))$, $o\bar{e}_j^t = E_t(\bar{f}_j^t)$, $o\bar{c}_j^t = C_t(\bar{f}_j^t)$.

The feature extraction module turns the target sample into features in a low-dimensional feature space. These features seem to have a clustering phenomenon [22], the parts belonging to the same class aggregated into a cluster, and samples corresponding to features closer to the centroid in each group are more likely to be correctly classified by the subsequent classifier. Assuming that the pseudo-labels corresponding to the parts closer to the centroid in each cluster are accurate, these features can represent the distribution of a particular type of sample in the feature space. This study refers to the samples corresponding to these features as HCS. Based on the above findings, this paper proposes a strategy that can accurately screen the above samples and is named high-confidence target domain sample acquisition strategy. As shown in Fig. 1, using three memory banks, we select the HCS through a joint confidence enhancement scheme based on the calculation of evidence theory, auxiliary correction with the entropy criterion, and feature distance.

3.3.1. High-confidence target domain sample acquisition

Evidence theory is a decision-making method that effectively expresses uncertain and unknown information, while the main difficulty is obtaining evidence. In this task, we refer to evidence as a measure of the amount of support collected from data that facilitates classifying a sample into a certain class. As shown in Fig. 1, we use OE bank to store the evidence vectors of sample $\bar{e}_j^t = \varphi(o\bar{e}_j^t)$. The uncertainty calculation for all target domain samples based on evidence theory is as follows:

$$UNC\_LIST = [unc_1^t, \ldots, unc_{n_t}^t], \tag{5}$$

where $unc_j^t = K / \sum_{k=1}^{K}(\varphi_k(o\bar{e}_j^t) + 1)$ is the uncertainty of sample x_j^t. Then we select the low uncertainty samples using

$$CI_{unc} = \{j \mid unc_j^t < \sigma_{unc}\}, \quad j \in [1, n_t], \tag{6}$$

where the threshold σ_{unc} is set to the median of the uncertainty over all samples.

Deterministic samples screened by evidence theory assist the network in improving the classification precision. To obtain more accurate samples close to the cluster centers, we would like to impose auxiliary correction through the information entropy criterion, and feature distance to the samples found based on evidence theory in mutual selection. Information entropy is used on the OC bank to compute the entropy of all target domain data as

$$ENT\_LIST = [ent_1^t, \ldots, ent_{n_t}^t], \tag{7}$$

where $ent_j^t = -\sum_{k=1}^{K} \delta_k(oc_j^t) \log \delta_k(oc_j^t)$ is the information entropy of sample x_j^t. Then we use the information entropy criterion to select low entropy samples defined by

$$CI_{ent} = \{j \mid ent_j^t < \sigma_{ent}\}, \quad j \in [1, n_t], \tag{8}$$

where the threshold σ_{ent} is set to the median of the entropy value over all samples.

Before implementing distance-based sample correction, we first calculate the pseudo centroid of the target sample μ to prepare for the subsequent calculation of feature distance, which is defined by

$$\mu = \{\mu_k\}_{k=1}^{K}, \quad \mu_k = \frac{\sum_{j=1}^{n_t} \delta_k(o\bar{c}_j^t)\bar{f}_j^t}{\sum_{j=1}^{n_t} \delta_k(o\bar{c}_j^t)}. \tag{9}$$

The distance between each sample feature and its corresponding nearest category centroid feature is defined by

$$DIS\_LIST = [dis_1^t, \ldots, dis_{n_t}^t], \tag{10}$$

where $dis_j^t = \min(1 - (\bar{f}_j^t)^T \cdot \mu_k / (\| \bar{f}_j^t \| \cdot \| \mu_k \|)), k \in [1, K]$ denotes the distance between sample feature $\bar{f}_j^t$ and its nearest class centroid feature μ_k. Then, the distance-based sample correction in a compact feature space is used to screen samples with a low distance defined by

$$CI_{dis} = \{j \mid dis_j^t < \sigma_{dis}\}, \quad j \in [1, n_t], \tag{11}$$

where the threshold σ_{dis} is the median of the distance. Finally, we conduct a joint confidence enhancement scheme on all these sets and obtain the final HCS set via $CI = CI_{ent} \cap CI_{unc} \cap CI_{dis}$.

3.3.2. Pseudo-labeling generation based on high-confidence samples

As shown in Fig. 1, instead of computing pseudo-labels using the features of all samples, this paper uses HCS to infer pseudo-label for each target domain instance in a self-supervised manner. Since HCS exhibits higher inter-class distinguishability and intra-class aggregation in feature space, the confidence category centroids μ^{CI} that are generated from HCS will have high accuracy defined as

$$\mu^{CI} = \{\mu_k^{CI}\}_{k=1}^{K}, \quad \mu_k^{CI} = \frac{\sum_{j \in CI} \delta_k(o\bar{c}_j^t)\bar{f}_j^t}{\sum_{j \in CI} \delta_k(o\bar{c}_j^t)}. \tag{12}$$

Subsequently, pseudo-labels for all samples are generated from these nearest centroids.

$$\tilde{Y}^t = \{\tilde{y}_j^t\}_{j=1}^{n_t}, \quad \tilde{y}_j^t = \arg\min_k \left(1 - \frac{(\bar{f}_j^t)^{\mathrm{T}} \cdot \mu_k^{CI}}{\| \bar{f}_j^t \| \cdot \| \mu_k^{CI} \|}\right). \tag{13}$$

The following pseudo-code indicates the HCS acquisition and pseudo-label generation implementation details.

3.3.3. Loss functions

The target model $\phi_t = \{\mathcal{F}_t, G_t, E_t, C_t\}$ consists of a feature extractor $\mathcal{F}_t$, bottleneck network G_t, evidence analysis E_t and classifier C_t, and the pre-trained source model initializes the parameters of each module. As shown in Fig. 1, in the target adaptation stage, when the target samples are used as input, they yield feature output $f_j^t = G_t(\mathcal{F}_t(x_j^t))$, evidence output $O_e^t = oe_j^t = E_t(f_j^t)$ and classification output $O_c^t = oc_j^t = C_t(f_j^t)$. The target model utilizes four types of loss functions to update the parameters.

$$\mathcal{L}_{tar} = \mathcal{L}_{im}^t + \alpha \mathcal{L}_{sl}^t + \beta \mathcal{L}_{edl}^t + \gamma \mathcal{L}_{ecl}^t, \tag{14}$$

/* HCS acquisition and pseudo-labels generation */
Input: *Feature* Bank F; *OE* Bank; *OC* Bank.
Output: HCS set CI; target sample pseudo-labels $\tilde{Y}^t$.
Compute unc_j^t and *UNC_LIST* using (5);
Compute ent_j^t and *ENT_LIST* using (7);
for each $j \in [1, n_t]$ **do**
    Compute $dis_j^t = \min(1 - (\bar{f}_j^t)^T \cdot \mu_k / (||\bar{f}_j^t|| \cdot ||\mu_k||))$;
end for
DIS_LIST $\leftarrow (dis_1^t, ..., dis_{n_t}^t)$;
Calculate CI_{unc}, CI_{ent}, and CI_{dis} using (6) (8), and (11), respectively;
Calculate HCS set: $CI = CI_{ent} \cap CI_{unc} \cap CI_{dis}$;
Compute high-confidence target pseudo-centroid μ^{CI} using (12);
for each $j \in [1, n_t]$ **do**
    Compute $\tilde{y}_j^t = \arg\min_k (1 - (\bar{f}_j^t)^T \cdot \mu_k^{CI} / (||\bar{f}_j^t|| \cdot ||\mu_k^{CI}||))$;
end for
Target sample pseudo-labels $\tilde{Y}^t \leftarrow (\tilde{y}_j^t, ..., \tilde{y}_{n_t}^t)$.

where γ, β, and α are three weights of neighborhood-guided evidence-based contrast loss $\mathcal{L}_{ecl}^t$, evidence loss $\mathcal{L}_{edl}^t$, and self-learning loss $\mathcal{L}_{sl}^t$, respectively. Using the HCS updated intermittent training scheme, the HCS, and pseudo-labels obtained during the HCS and pseudo-labels update make one obtain more reliable self-supervised loss and neighborhood-guided evidence-based contrast loss in the model update process. Information maximization (IM) loss $\mathcal{L}_{im}^t$ guides global clustering, and evidence loss promotes the utilization of network output information.

Information maximization loss. Generally, the desired outputs after the target domain data passes through the network are similar to one-hot encoding. We know that minimizing IM loss corresponds to maximizing mutual information. Therefore, this paper adopts IM loss to make the target output individually determined and globally diversified. IM consists of conditional entropy minimization and target diversification terms formulated by

$$\mathcal{L}_{im}^t = -\frac{1}{n_t}\sum_{j=1}^{n_t}\sum_{k=1}^{K}\delta_k(oc_j^t)\log\delta_k(oc_j^t) + \sum_{k=1}^{K}\bar{p}_k\log\bar{p}_k, \tag{15}$$

where $\bar{p}_k = E_{j\in[1,n_t]}(\delta(oc_j^t))$ is the mean of the Softmax outputs. The first term ensures global clustering, while the second term balances cluster allocations and encourages diversity in cluster aggregation.

Self-supervised loss. Once a reliable pseudo-label is acquired for each sample with the method mentioned above, the model parameters can be updated in a self-supervised manner. At this time, the self-supervised loss that enforces semantic alignment is

$$\mathcal{L}_{sl}^t = -\frac{1}{n_t}\sum_{j=1}^{n_t}\sum_{k=1}^{K}\tilde{l}_{j,k}^t\log\delta_k(oc_j^t), \tag{16}$$

where $\tilde{l}_j^t$ is a one-hot encoding [2] of the target domain pseudo-label $\tilde{y}_j^t$, and $\tilde{l}_{j,k}^t$ is the kth element of $\tilde{l}_j^t$.

Evidence loss of target adaptation. To guide the proposed UC-SFDA to make accurate predictions over the target domain, we add an evidence loss to optimize the network defined by

$$\begin{aligned}\mathcal{L}_{edl}^t = &\sum_{j=1}^{n_t}\sum_{k=1}^{K}\mathbb{E}[(\tilde{l}_{j,k}^t - \varphi_k(oe_j^t))^2] \\ &+ \lambda_t\sum_{j=1}^{n_t}KL[D(\varphi(oe_j^t)\mid\theta_j^t)\;||\;D(\varphi(oe_j^t)\mid\mathbf{I})],\end{aligned} \tag{17}$$

where the parameter of the Dirichlet distribution is $\theta_j^t = \varphi_k(oe_j^t)+1$. With the aid of evidence loss, the designed network can provide a more detailed estimation of the probability distribution so that the network can utilize the output information of both the evidence analyzer and classifier more comprehensively.

Neighborhood-guided evidence-based contrast loss. Contrastive learning aims to achieve instance-level recognition by keeping separate semantically different points and close semantically similar points in the feature space. For an "anchor" data x, our goal is to train an encoder f to make $score(f(x), f(x^+)) >> score(f(x), f(x^-))$, where a positive pair x^+ refers to data similar to x, while a negative pair x^- refers to data different from x. The positive and negative pairs are named "keys". The *score* is used to compare the similarity between the two features. By observing that the feature distribution of intra-class target domain data is more likely to be distributed in the neighboring area, we believe the neighborhood data structure helps solve SFDA. In addition, choosing positive and negative pairs based on HCS and corresponding pseudo-labels can reduce the influence of uncertain noise samples. Therefore, this study proposes a neighborhood-guided evidence-based contrast loss defined as follows:

$$\mathcal{L}_{ecl}^t = \frac{Dis(q_j, q_j^{neb}) + Dis(q_j, key_j^+)}{Dis(q_j, key_j^+) + Dis(q_j, key_j^-)}, \tag{18}$$

where $Dis(\cdot,\cdot)$ measures feature distances, we take the feature vector f_j^t as anchor vector q_j, q_j^{neb} represents the high-confidence neighbor features of the anchor, and the number is n_{neb}. The positive pairs key_j^+ are features attributed to the same group as q_j, and we set the number of positive pairs to 1. In contrast, the negative pair key_j^- is the feature belonging to a different category from q_j. The number of negative pairs is $n_{neg} = K - 1$. Thus, all pairs are $key_j = key_j^+ \cup key_j^-$.

Next, we use the HCS set CI to find the corresponding target domain feature set $F_{con} = \{\bar{f}_j^t\}, j \in CI$, and then the high-confidence neighbor features q_j^{neb} of q_j can be obtained by the following search method.

$$q_j^{neb} = f_j^{neb} = Top_{F_{con}}(n_{neb}, 1/Dis(q_j, F_{con})), \tag{19}$$

where $Top_{F_{con}}(n_{neb}, 1)/Dis(q_j, F_{con})$ is a set composed of the first n_{neb} elements in F_{con} that maximize the value of $1/Dis(q_j, F_{con})$. Since the high-confidence target domain features are selected via evidence theory and further corrected by the entropy criterion and feature distance, the pseudo-labels corresponding to F_{con} have high accuracy, so we use high-confidence category centroids μ^{CI} as the memory bank for the anchor to select positive and negative pairs. The centroid features in μ^{CI} belonging to the same class as q_j are selected as the positive pairs key_j^+ and are obtained by

$$key_j^+ = Top_{\mu^{CI}}(1, 1/Dis(q_j, \mu^{CI})), \tag{20}$$

and the negative pairs key_j^- are the centroid features in μ^{CI} except for key_j^+, which belong to different categories from q_j. Next, we rewrite the neighborhood-guided evidence-based contrast loss using the dot product as follows:

$$\mathcal{L}_{ecl}^t = -\frac{1}{n_t}\sum_{j=1}^{n_t}\log\frac{\exp(q_j\cdot key_j^+/\tau) + \exp(q_j\cdot q_j^{neb}/\tau)}{\exp(q_j\cdot key_j^+/\tau) + \exp(q_j\cdot key_j^-/\tau)}, \tag{21}$$

where τ is a trade-off weight, and we take the input sample feature together with their high-confidence neighbor features as queues. In this way, multiple semantically similar points in the feature space gather together to improve the compactness of the class. We choose category centroids as negative and positive pairs. The advantage is that samples of the same category move

toward the centroids of such instances without biasing any single sample while semantically pushing away different points and their neighbors relative to different class centroids.

We use batch training to fine-tune all trainable parameters in the actual training process. To better understand the optimization process, we present the implementation details of $\mathcal{L}_{tar}$ with the following pseudo-code.

/* Implementation details of $\mathcal{L}_{tar}$ during batch target adaptation */
Input: Dataset $\mathcal{D}_t$; model ϕ_t; *Feature* bank F; HCS set CI; target sample pseudo-labels $\tilde{Y}^t$; high-confidence category centroids μ^{CI}; max training epoch $T_{\max}$; one batch of target data X_{Batch}.
Output: Loss function $\mathcal{L}_{tar}$.
Compute one-hot encoding $\tilde{l}_j^t$ in terms of [2];
for each idx $\in CI$ **do**
  add HCS features $\tilde{f}_{idx}^t$ to target domain feature set F_{con};
end for
for each epoch $\in [1, T_{\max}]$ **do**
  for each $x_n^t \in X_{Batch}$ **do**
    Compute sample feature q_n, evidence output oe_n^t;
    Compute classifier output oc_n^t;
    Compute mean value of the Softmax output $\bar{p}_k$;
    Calculate $\mathcal{L}_{im}^t$, $\mathcal{L}_{sl}^t$ and $\mathcal{L}_{edl}^t$ using (15), (16) and (17);
    Compute HCS neighbor features q_n^{neb} and positive pairs key_n^+ using (19) and (20);
    Compute key_n^- according to $key_n^- \cup key_n^+ = \mu^{CI}$;
    Compute $\mathcal{L}_{ecl}^t$ and $\mathcal{L}_{tar}$ using (21) and (14);
  end for
end for

4. Results and discussion

4.1. Datasets

The Office-31 dataset [23] is the most classic real-world domain adaptation dataset. It is an unbalanced dataset containing 4110 images in 31 categories collected from three domains: DSLR (D), Webcam (W), and Amazon (A). Set D contains 795 images taken by digital SLR cameras. Set W contains 498 images taken by a webcam. Set A includes 2817 images downloaded from Amazon's official website.

The Office–Home dataset [24] contains 15,500 images of 65 categories taken in offices and homes. It consists of images of everyday items divided into four domains: 4269 real-world images (Rw), 4439 product images without background (Pr), 4365 clipart images (Cl), and 2427 art images (Ar).

4.2. Implementation details

The proposed UC-SFDA is programmed in PyTorch 1.9.0 and trained on a server with Intel(R) Core(TM) i9-12900K CPU and NVIDIA GeForce RTX3090 GPU. The pre-trained ResNet-50 serves as a base feature extractor of dimension 2048, followed by a bottleneck network composed of FC layers (2048×256) and BN layers. The evidence analyzer and classifier are set in parallel. Both consist of an FC layer ($256 \times K$), followed by ReLU and softmax activation. ResNet-50 with transformer layer (see Fig. 1) is initialized using the pre-trained R50+ViT-B_16 model on the ImageNet-21K dataset in ViT [25]. The study uses the stochastic gradient descent optimizer with a momentum of 0.9 and a batch of 64 to fine-tune all trainable parameters. The initial learning rate η_0 of the feature extractor is set to 1×10^{-2}, and for the bottleneck network, evidence analyzer and classifier, the initial learning rate η_0 is defined as 1×10^{-3}. The learning rate follows the same rule in [26], that is $\eta = \eta_0 \cdot (1 + 10 \cdot p)^{-0.75}$, where p changes linearly from 0 to 1. Referring to [8,14], the weights

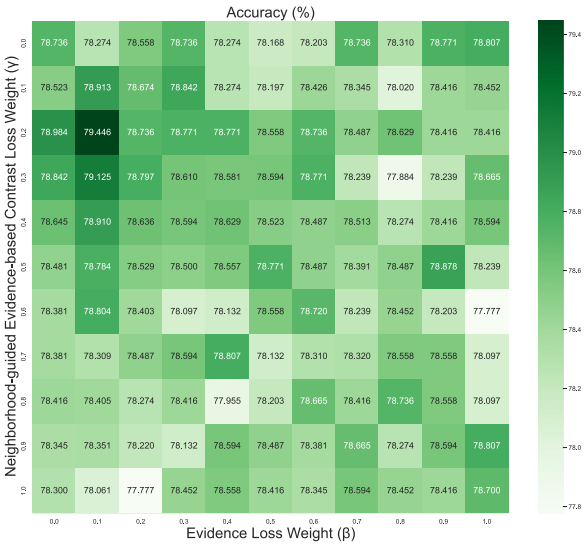


Fig. 2. Evaluation of weights β and γ for D→A task on Office-31. They are used to balance evidence loss and neighborhood-guided evidence-based contrast loss.

τ and α are set as 0.07 and 0.03 respectively. The following experiment will discuss the impact of the other weights n_{neb}, β, and γ.

4.3. Discussion of trade-off weights

This experiment first investigates the impact of weights β and γ in (14), where the former controls the contributions of the evidence loss, and the latter balances the effect of neighborhood-guided evidence-based contrast loss. In this experiment, the optimal weights are selected by observing the target domain classification accuracy on Office-31. Concretely, we adjust β and γ from 0 to 1 in 0.1 intervals, and the classification accuracy matrix on the D→A task is illustrated in Fig. 2. We can see that both weights will slightly affect the classification performance, while $\beta = 0.1$ and $\gamma = 0.2$ maintain at the highest results (79.446%).

Next, we analyze the effect of different numbers of n_{neb} on classification performance. For this, we conduct experiments with n_{neb} varying from 0 to 5 with a step size of 1, and the $\mathcal{L}_{ecl}^t$ curve and classification accuracy of the Pr→Ar task on Office–Home are reported in Fig. 3. The results show that the best result occurs when $n_{neb} = 1$. For the classification accuracy experiment, it is observed from the first half of the training that the proposed model demonstrates a similar accuracy curve as n_{neb} gradually increases from 0 to 2. The optimal performance is achieved when setting $n_{neb} = 1$. This phenomenon can be explained as follows: the distance between the sample features q and their neighbor features q_{neb} increases, and the similarity between q and q_{neb} decreases. Therefore, as n_{neb} continues to increase, more neighbors q_{neb} of feature q will be regarded as the same category, reducing the classification precision.

4.4. Performance comparison

We consider image classification to evaluate the proposed method on major domain adaptation benchmarks, including (i) source-only, target prediction by our network using a pre-trained source model; (ii) UDA method, the adaptation process can access the source domain including MCD [27], CDAN [28], MDD [29], CAN [30], DMRL [31], BDG [32], MCC [33], SRDC [34],

Table 1
Comparison of classification accuracy (%) on Office-31 (ResNet-50). The red/blue texts denote the first/second best results. SF denotes source-free. * represents backbone network with transformer layer, Ours-I denotes our UC-SFDA with ResNet-50 and without transformer layer.

Types	Methods	SF	D→A	A→D	W→A	A→W	D→W	W→D	Avg
UDA	MCD (CVPR'18)	✗	69.5	92.2	69.7	88.6	98.5	100.0	86.5
	CDAN (NeruIPS'18)	✗	71.0	92.9	69.3	94.1	98.6	100.0	87.7
	MDD (ICML'19)	✗	75.0	90.4	73.7	90.4	98.7	99.9	88.0
	CAN (CVPR'19)	✗	70.3	95.0	66.4	94.5	99.1	99.6	90.6
	DMRL (ECCV'20)	✗	73.0	93.4	71.2	90.8	99.0	100.0	87.9
	BDG (AAAI'20)	✗	73.2	93.6	72.0	93.6	99.0	100.0	88.5
	MCC (ECCV'20)	✗	72.6	95.6	73.9	95.4	98.6	100.0	89.4
	SRDC (CVPR'20)	✗	76.7	95.8	77.1	95.7	99.2	100.0	90.8
	BNM (CVPR'20)	✗	70.9	90.3	71.6	91.5	98.5	100.0	87.1
	RWOT (CVPR'20)	✗	77.5	94.5	77.9	95.1	99.5	100.0	90.8
	RSDA-MSTN (CVPR'20)	✗	77.4	95.8	78.9	96.1	99.3	100.0	91.1
Source	Source only	✗	62.1	80.7	65.1	75.5	96.1	97.9	79.5
	Source only*	✗	63.2	88.2	66.1	88.3	98.4	99.8	84.0
SFDA	SFDA (arXiv'20)	✗	71.0	92.2	71.2	91.1	98.2	99.5	87.2
	SHOT (ICML'20)	✓	74.7	94.0	74.3	90.1	98.4	99.9	88.6
	MBNS (arXiv'21)	✓	78.5	89.0	76.6	91.7	98.9	100.0	89.1
	BAIT (arXiv'20)	✓	74.6	92.0	75.2	94.6	98.1	100.0	89.1
	NRC (NeruIPS'21)	✓	75.3	96.0	75.0	90.8	99.0	100.0	89.4
	3C-GAN (CVPR'20)	✓	75.3	92.7	77.8	93.7	98.5	99.8	89.6
	VDM-DA (TCSVT'22)	✓	75.8	93.2	77.1	94.1	98.0	100.0	89.7
	HCL (NeruIPS'21)	✓	75.9	94.7	77.7	92.5	98.2	100.0	89.8
	CPGA (arXiv'21)	✓	76.0	94.4	76.6	94.1	98.4	99.8	89.9
	LPA-SFDA (arXiv'22)	✓	75.8	94.8	76.8	93.7	98.4	100.0	89.9
	JN-USFDA (arXiv'22)	✓	75.4	95.9	77.4	92.5	98.5	99.9	89.9
	N2DCEX (arXiv'21)	✓	75.4	97.0	75.6	93.0	98.9	99.8	90.0
	SFDA-DE (arXiv'22)	✓	76.6	96.0	75.5	94.2	98.5	99.8	90.1
	Ours-I	✓	76.4	96.0	74.6	95.4	99.0	100.0	90.2
	TransDA* (arXiv'21)	✓	73.7	97.2	79.3	95.0	99.3	99.6	90.7
	Ours*	✓	79.5	96.6	80.5	96.2	99.3	100.0	92.0

Table 2
Comparison of classification accuracy (%) on Office–Home (ResNet-50). The red/blue texts denote the first/second best results. SF denotes source-free. * represents backbone network with transformer layer.

Types	Methods	SF	Cl→Ar	Pr→Ar	Rw→Ar	Ar→Cl	Pr→Cl	Rw→Cl	Ar→Pr	Cl→Pr	Rw→Pr	Ar→Rw	Cl→Rw	Pr→Rw	Avg
UDA	MCD (CVPR'18)	✗	61.3	57.0	69.1	48.9	47.1	52.2	68.3	67.6	79.6	74.6	68.8	75.1	64.1
	CDAN (NeruIPS'18)	✗	57.6	57.4	70.9	50.7	50.9	56.7	70.6	70.0	81.6	76.0	70.0	77.3	65.8
	SAFN (ICCV'19)	✗	64.2	63.7	70.9	52.0	51.4	57.1	71.7	69.9	81.5	76.3	71.9	77.1	67.3
	Symnets (CVPR'19)	✗	64.2	64.2	74.5	47.7	48.8	52.6	72.9	71.3	82.7	78.5	74.2	79.5	67.6
	TADA (AAAI'19)	✗	59.1	59.7	72.4	53.1	53.1	60.0	72.3	71.2	82.9	77.2	72.1	78.4	67.6
	BNM (CVPR'20)	✗	63.3	61.7	70.5	52.3	49.5	53.6	73.9	72.9	82.2	80.0	74.9	79.7	67.9
	MDD (ICML'19)	✗	60.0	61.2	72.5	54.9	53.6	60.2	73.7	71.4	82.3	77.8	71.8	78.1	68.1
	BDG (AAAI'20)	✗	65.3	65.1	74.6	51.5	49.7	55.1	73.4	71.5	84.8	78.7	73.7	81.1	68.7
	SRDC (CVPR'20)	✗	69.5	68.7	76.3	52.3	53.8	57.1	76.3	76.2	85.0	81.0	78.0	81.7	71.3
Source	Source only	✗	54.3	53.2	66.2	44.6	41.8	46.6	65.5	61.8	77.8	74.4	64.8	73.3	60.3
	Source only*	✗	61.9	69.6	75.9	58.6	58.1	61.2	73.7	69.2	84.1	80.1	72.5	83.4	70.7
SFDA	SFDA (arXiv'20)	✓	64.3	62.7	69.8	48.4	45.3	50.5	73.4	69.8	79.0	76.9	71.7	76.6	65.7
	CPGA (arXiv'21)	✓	65.4	65.7	72.0	59.3	58.0	64.4	78.1	75.5	83.3	79.8	76.4	81.0	71.6
	BAIT (arXiv'20)	✓	68.0	67.1	73.9	57.4	55.5	59.5	77.5	77.2	84.2	82.4	75.1	81.9	71.6
	SHOT (ICML'20)	✓	68.0	67.4	73.3	57.1	54.9	58.8	78.1	78.2	84.3	81.5	78.1	82.2	71.8
	PS (arXiv'21)	✓	68.4	67.8	75.2	57.8	57.3	59.1	77.3	76.9	83.4	81.2	78.1	82.1	72.1
	NRC (NeruIPS'21)	✓	68.1	65.3	71.0	57.7	56.4	58.6	80.3	79.8	85.6	82.0	78.6	83.0	72.2
	LPA-SFDA (arXiv'22)	✓	68.9	67.2	72.1	59.3	57.4	58.5	79.3	79.8	85.4	82.1	79.5	83.1	72.7
	SFDA-DE (arXiv'22)	✓	69.7	66.1	73.9	59.7	57.2	60.8	79.5	78.6	85.5	82.4	79.2	82.6	72.9
	N2DCEX (arXiv'21)	✓	69.8	68.7	74.4	57.4	56.5	60.4	80.0	79.6	85.6	82.1	80.3	82.6	73.1
	JN-USFDA (arXiv'22)	✓	68.3	68.8	75.3	56.4	55.1	58.8	78.6	80.1	84.6	82.0	79.4	82.2	74.5
	Ours-I	✓	69.9	69.8	74.8	58.2	56.6	60.0	80.1	81.6	85.6	81.7	80.2	82.4	73.4
	TransDA* (arXiv'21)	✓	74.0	77.0	80.5	67.5	68.0	69.9	83.3	83.8	90.0	85.9	84.4	87.0	79.3
	Ours*	✓	76.9	79.1	81.6	69.1	69.6	70.7	84.9	85.3	88.4	86.5	84.4	86.7	80.3

BNM [35], RWOT [36], RSDA-MSTN [37], TADA [38], SAFN [39], and Symnets [40]; and (iii) SFDA method, the adaptation process cannot access the source domain including SFDA [41], SHOT [2], MBNS [42], BAIT [43], NRC [10], 3C-GAN [1], PS [5], VDM-DA [7], HCL [9], CPGA [6], N2DCEX [3], TransDA* [8], LPA-SFDA [44], SFDA-DE [45], and JN-USFDA [46]. Tables 1 and 2 report the classification precision on the Office-31 and Office–Home datasets, respectively. For a fair comparison, we first used ResNet-50 pre-trained on ImageNet as the backbone; simultaneously, we removed the "transformer layer" from our framework, and the

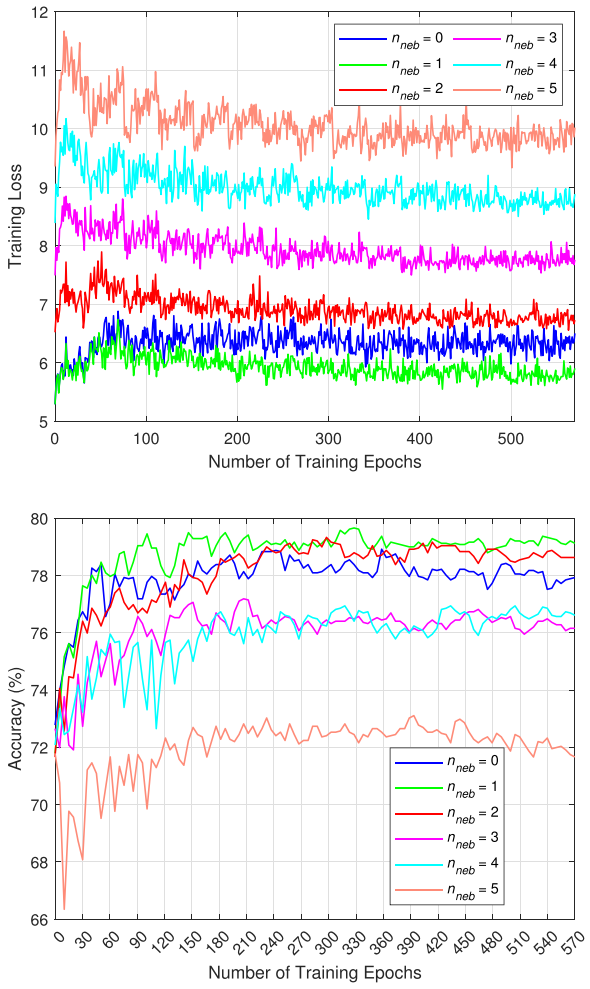


Fig. 3. Neighborhood-guided evidence-based contrast loss and classification accuracy under different numbers of neighbors (Pr→Ar task on Office–Home).

result is denoted as Ours-I. * represents the backbone network with the transformer layer. Specifically comparing our performance with or without the transformer layer, Ours-I achieves the best average classification performance compared to the SFDA algorithm with the same backbone network ResNet-50. This explicitly shows the competitiveness of our network apart from applying the transformer structure due to the evidence theory adopted for uncertainty prediction throughout the network. With the transformer layer, the proposed UC-SFDA not only surpasses all UDA algorithms with ResNet-50 as the backbone network, but also outperforms TransDA, which is the latest method with a transformer layer in 6 out of 7 subtasks. For the two most difficult tasks, the W→A and D→A Office-31 dataset, our model's accuracies reach 79.5% and 80.5%, which are 16.3% and 14.4% higher than that of the source-only method, respectively. This difference indicates that uncertain prediction using evidence theory and the transformer layer could improve the classification performance.

On the Office–Home dataset, our method achieves an average accuracy of 80.3% (shown in the last column of Table 2), which is not inferior to other advanced SFDA algorithms, whether they need to use the source data or not. Comparing the last two rows in Table 2, one can observe that TransDA* outperformed our method on 1 out of 13 subtasks. Both the proposed method and TransDA* benefit from the self-attention mechanism of the transformer layer, so they greatly outperform the other algorithms. The difference between our approach and TransDA* is that the latter uses self-supervised pseudo-labeling and knowledge distillation schemes to improve the network's ability to pay attention to a target object. Our approach uses evidence theory combined with contrastive learning to further facilitate inter-class separability and intra-class compactness. Thus the average accuracy on both datasets exceeds TransDA*. Overall, uncertainty prediction using evidence theory and the HCS updated intermittent training scheme is helpful for SFDA tasks.

4.5. Network complexity and time consumption analysis

In this section, we compare the network parameters and time consumption of the proposed UC-SFDA with the baseline methods under two different backbone networks (w/ transformer layer, w/o transformer layer). The proposed algorithm w/ transformer layer is denoted as Ours, and TransDA is its corresponding baseline method. The proposed algorithm w/o transformer layer is denoted as Ours-I, and SHOT is its corresponding baseline method.

As shown in Figs. 4 (a) and (d), the size of the proposed model is almost the same as that of the corresponding baseline method. The largest increase in model parameters is 0.065 MB on Office–Home w/ transformer, a relative increase of less than 0.17%, which is considered negligible. However, the classification performance of the proposed model is significantly improved compared with the corresponding backbone network (1% ∼ 1.6%).

Since the SFDA algorithm is divided into source training and target adaptation stages, we compared the time consumption of the proposed algorithm and the baseline methods in both stages. Figs. 4 (b) and (e) show that the time consumption of the proposed algorithm for source training on each task of Office-31 is slightly greater than that of the baseline method. Specifically, the average time increases for Ours-I compared to SHOT and Ours compared to TransDA are 2.12 min and 1.95 min, respectively. Our evidence analyzer structure and additional evidence loss of the proposed model result in extra training time. However, sufficient improvement in performance is achieved. The average classification accuracy for Ours-I compared to SHOT and Ours compared to TransDA increased by 0.3% and 2.1%, respectively.

In addition, for Figs. 4 (c) and (f), the time for Ours-I and Ours to complete the target adaptation on each task of Office-31 is greater than that of the corresponding backbone networks SHOT and TransDA and the average adaptation time increased by 4 min and 4.9 min, resulting in an average accuracy increase of 1.6% and 1.3%, respectively. This phenomenon can be explained by introducing the HCS update intermittent training scheme into the proposed UC-SFDA. In the HCS and pseudo-labels update process, it is necessary to prepare the data needed for the model update process, so extra time is needed. However, introducing this strategy improves SFDA performance, and even on the most challenging D→A task, ours improves classification accuracy by up to 5.8% compared to the baseline network TransDA. In summary, although the proposed algorithm has higher time consumption compared to the baseline network, the final performance is sufficiently improves.

4.6. Performance analysis

Visualization analysis using t-SNE. This experiment uses t-SNE [47] to visualize the target domain features for two subtasks D→A and A→D on Office-31. As shown in Fig. 5, UC-SFDA can encourage the model to align cross-domain features better, indicating that UC-SFDA produces more discriminative features. In Fig. 5, we use t-SNE to project the last hidden layer features onto the 2D space and show the t-SNE results before and after domain adaptation in the tasks of Pr→Ar and Ar→Pr on Office–Home. Here different colors represent different categories. The

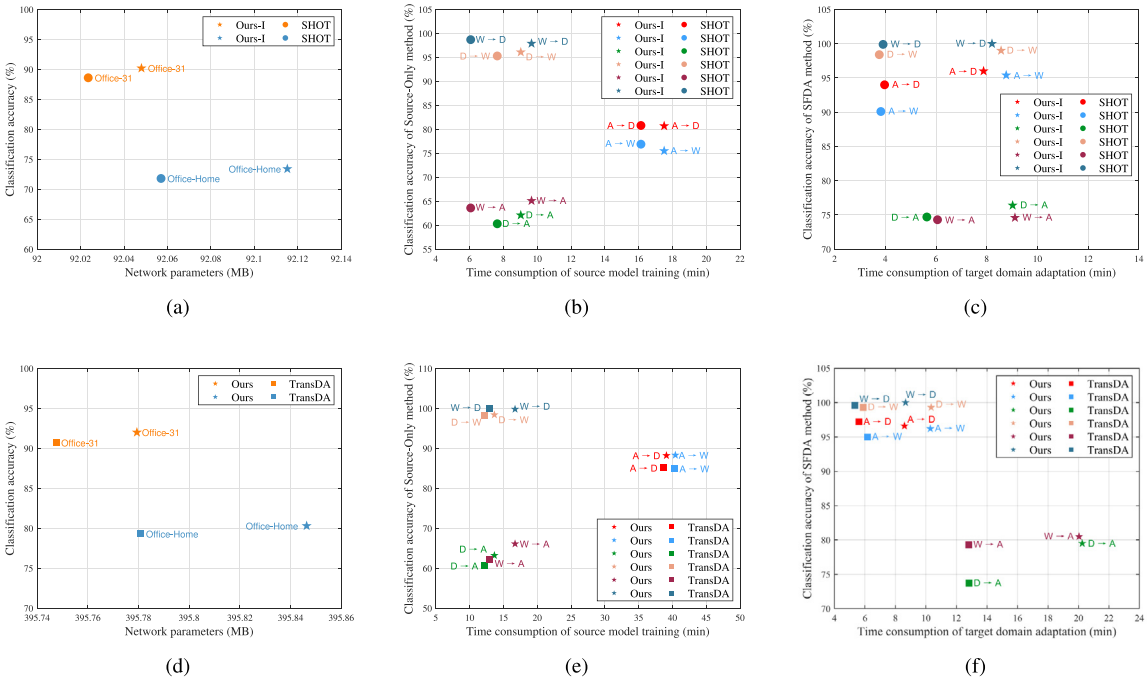


Fig. 4. Comparison results of network parameters and time consumption between the proposed algorithm and the baseline methods under two backbone networks, without (w/o) transformer layer (Row 1) and with (w/) transformer layer (Row 2). Column 1 shows the network parameters, and Columns 2 and 3 show the time consumption of source training and target adaptation.

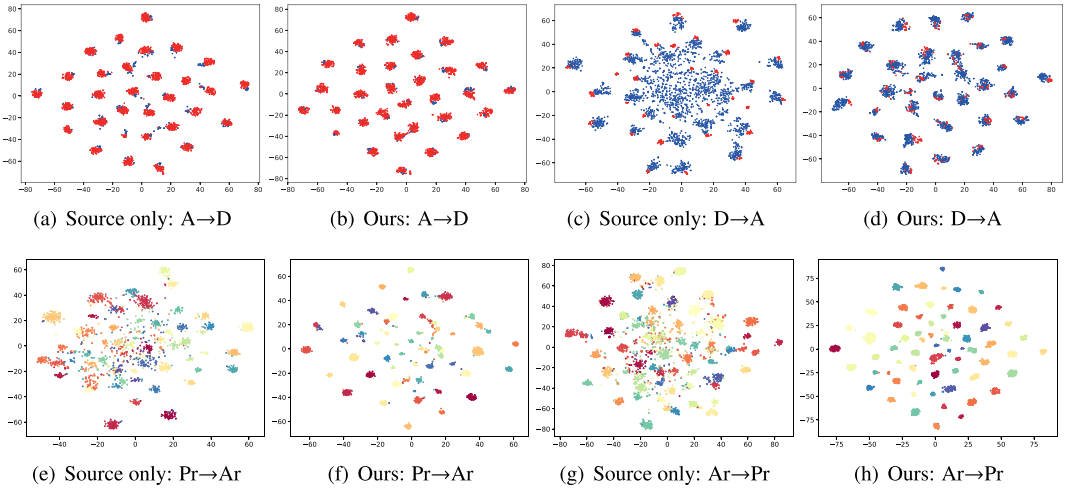


Fig. 5. The first row is feature space visualization on Office-31. Red and blue denotes the source and target domains, respectively. The second row is t-SNE visualizations for two classification tasks on Office–Home. Categories are indicated in different colors.

target samples are obviously clustered for each category. Our method achieves more obvious classification decision boundaries between different classes.

Confusion matrix. We investigate the target domain classification results using the confusion matrix of the pre-trained source model and the proposed UC-SFDA for A→D and D→A cases on Office-31, as shown in Fig. 6, in which the x-axis denotes the predicted labels and the y-axis represents the ground truth labels. The results reveal that UC-SFDA results in significantly fewer misclassifications after domain adaptation.

Convergence. We then take the W→A task on Office-31 as an example to investigate the convergence of UC-SFDA in target domain adaptation by visualizing the curves of $\mathcal{L}_{im}^{t}$, $\mathcal{L}_{sl}^{t}$, $\mathcal{L}_{edl}^{t}$, and $\mathcal{L}_{ecl}^{t}$. As shown in Fig. 7, all loss functions converge gradually as the number of epochs increases. The classification accuracy of the target domain is gradually improved and finally stabilizes. The decreases of $\mathcal{L}_{edl}^{t}$ proves that the variance and prediction error of the Dirichlet distribution produced by the proposed evidence analyzer gradually converges. From loss function $\mathcal{L}_{sl}^{t}$, we observe that the semantic alignment of samples of the same class in the target domain gradually strengthens. The curve of $\mathcal{L}_{im}^{t}$ also indicates that our UC-SFDA can ensure the global clustering of target data. In addition, we find that the proposed $\mathcal{L}_{ecl}^{t}$ rapidly decreases during the training process, which indicates that it

Table 3
Ablation study of different network structures for several subtasks on Office-31 using Grad_cam++ and Heatmap visualization.

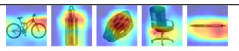
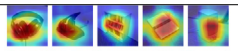
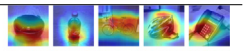
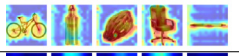
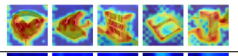
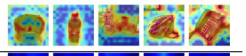
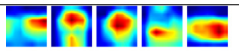
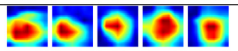
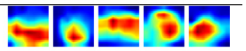
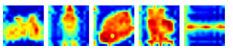
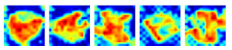
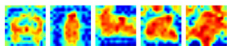
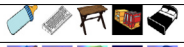
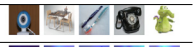
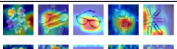
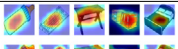
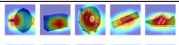
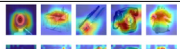
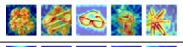
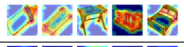
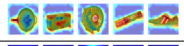
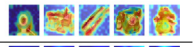
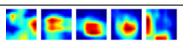
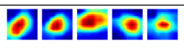
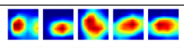
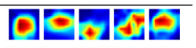
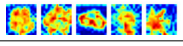
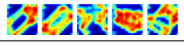
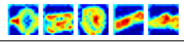
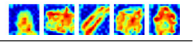
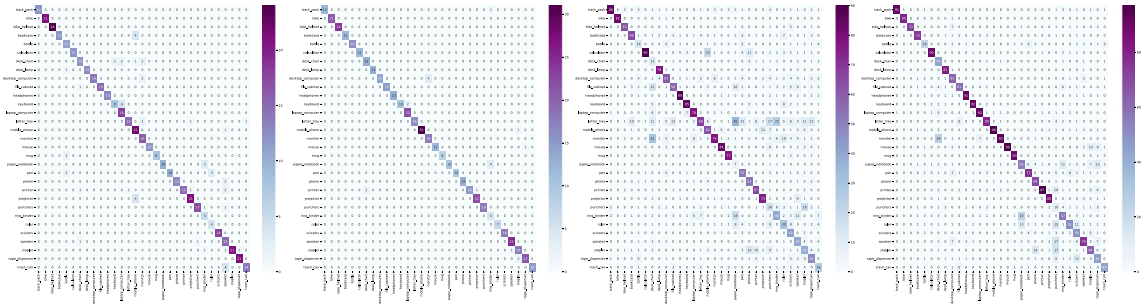
Tasks		W→A	A→D	D→W
Original Images				
Grad_cam++	w/o T_s+T_t			
	w/o E_s+E_t			
	Ours			
Heatmap	w/o T_s+T_t			
	w/o E_s+E_t			
	Ours			

Table 4
Ablation study of different network structures for several subtasks on Office–Home using Grad_cam++ and Heatmap visualization.

Tasks		Rw→Ar	Ar→Cl	Cl→Pr	Pr→Rw
Original Images					
Grad_cam++	w/o T_s+T_t				
	w/o E_s+E_t				
	Ours				
Heatmap	w/o T_s+T_t				
	w/o E_s+E_t				
	Ours				



(a) Source only: A→D (b) Ours: A→D (c) Source only: D→A (d) Ours: D→A

Fig. 6. Confusion matrix for A→D and D→A on Office-31 dataset, the *x*-axis shows the predicted labels, while the *y*-axis shows the ground truth.

makes good use of the inherent neighborhood information of the feature space and implements a neighborhood-guided evidence-based contrastive learning scheme. This enhances the inter-class separability and intra-class compactness of the target data in the feature space.

4.7. Ablation study

Effect of network structure. For the proposed model, transformer layer T_s is added to improve the generalization capability of the feature extraction model, and the evidence analyzer E_s is designed to operate a comprehensive probability estimate. Two improvements help to obtain a better pre-trained source model and improve accuracy in the target domain. Here, we perform ablation experiments on the network structure with (w/) and without (w/o) T_s and E_s. First, to intuitively show the advantages of our source training model over the public network, we compare the visualization results of Grad_cam++ [48] and Heatmap in Tables 3 and 4. Grad_cam++ uses gradients to find model attention representations. Thus, we can understand the decisions of any learning-based model. The visualization results of Grad_cam++ show the areas of interest of the model in red. Heatmaps represent salient features highlighted by the network. From Tables 3 and 4, one can see that when the transformer layer and evidence analyzer are added to the network, the red area on the target becomes more prominent and accurate. In contrast, the color of the background area is bluer, indicating that the network

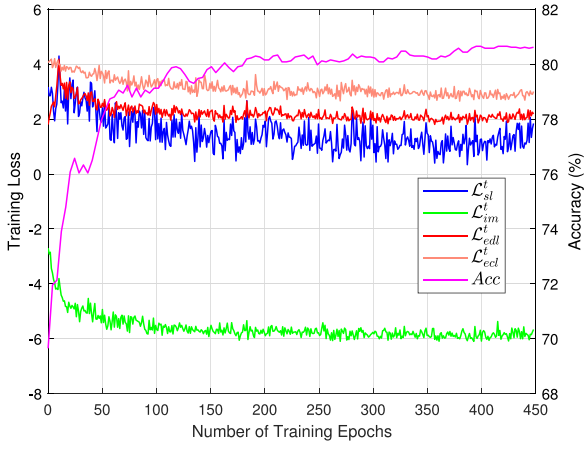


Fig. 7. Convergence analysis of the proposed UC-SFDA, including target domain classification accuracy and the trend of training loss for W→A task on Office-31.

Table 5
The classification accuracy of different network structures and losses on Office-31.

Network structures	D→A	A→D	W→A	A→W	D→W	W→D	Avg
w/o T_s+T_t , w/o E_s+E_t	75.5	95.3	75.1	91.4	98.9	100.0	89.4
w/ T_s+T_t , w/o E_s+E_t	79.3	96.3	79.7	96.0	99.1	100.0	91.7
w/o T_s+T_t , w/ E_s+E_t	76.4	96.0	74.6	95.4	99.0	100.0	90.2
Ours	79.5	96.6	80.5	96.2	99.3	100.0	92.0
Loss functions	D→A	A→D	W→A	A→W	D→W	W→D	Avg
$\mathcal{L}_{im}^t$ + $\mathcal{L}_{sl}^t$	78.5	96.1	78.5	94.7	99.1	100.0	91.2
$\mathcal{L}_{im}^t$ + $\mathcal{L}_{sl}^t$ + $\mathcal{L}_{ecl}^t$	78.9	96.4	79.4	94.9	99.2	100.0	91.5
$\mathcal{L}_{im}^t$ + $\mathcal{L}_{sl}^t$ + $\mathcal{L}_{edl}^t$	78.7	96.3	79.7	94.9	99.1	100.0	91.4
$\mathcal{L}_{im}^t$ + $\mathcal{L}_{sl}^t$ + $\mathcal{L}_{edl}^t$ + $\mathcal{L}_{ecl}^t$	79.5	96.6	80.5	96.2	99.3	100.0	92.0

is encouraged to pay more attention to the object. Overall, the visualization results of both Grad_cam++ and Heatmap show that the transformer layer and evidence analyzer play a crucial role in improving the model's generalization ability.

Table 5 summarizes the corresponding classification accuracies. It is observed that the transformer layer T_s significantly improves the generalization ability and achieves better classification results. On average, 1.8% decreases when ablating the transformer layer on Office-31. In addition, the evidence analyzer in the proposed model also increases the average accuracy by 0.2%. This shows that the proposed UC-SFDA with an evidence analyzer conducting detailed probability distribution estimates also improves classification precision. Overall, using the transformer layer and the evidence analyzer simultaneously achieves the optimal performance.

Contribution of loss function. We next explore the contribution of evidence loss $\mathcal{L}_{edl}^t$ and neighborhood-guided evidence-based contrast loss $\mathcal{L}_{ecl}^t$. The quantitative results of the proposed UC-SFDA optimized with different loss functions are reported in Table 5. As shown in this table, training on neighborhood-guided evidence-based contrast loss combination $\mathcal{L}_{im}^t$+$\mathcal{L}_{sl}^t$+$\mathcal{L}_{ecl}^t$ improves the performance on the target domain to some extent (on average 91.6% on Office-31). This result indicates that the neighborhood-guided evidence-based contrastive training scheme largely benefits target domain adaptation by enhancing the distinguishability between categories. Ablating $\mathcal{L}_{ecl}^t$ and using the combination $\mathcal{L}_{im}^t$+ $\mathcal{L}_{sl}^t$ + $\mathcal{L}_{edl}^t$ also boosts classification precision by achieving 91.5%. With the aid of evidence loss, the evidence analyzer provides more detailed estimates of probability distributions and comprehensively utilizes the output information. Overall, adding two losses together brings the most increase in classification accuracy from 91.2% to 92%.

Effect of the HCS selection strategy. To verify whether the proposed HCS acquisition strategy selects the correct data that can accurately describe the data distribution of the target domain, we perform ablation studies to investigate different selection strategies and summarize the classification results in Fig. 8. Here, *Unc*, *Ent*, and *Dis* represent the selection strategies based on evidence theory, entropy criterion, and feature space distance, respectively. *Null* represents no proposed HCS selection. Using all samples as HCS to calculate pseudo-labels shows very low accuracy. In contrast, applying all three HCS selection strategies brings an 18.73% improvement in the accuracy of HCS compared to the case of *Null*. Using a single HCS acquisition obtained a similar level of increase. Although the accuracy has already reached a high level, applying three HCS acquisition strategies improves the target domain classification precision (84.91%). We conclude that using evidence theory, entropy criterion, and feature space distance together and conducting a joint confidence enhancement scheme can yield the highest quality set of HCS. This also proves

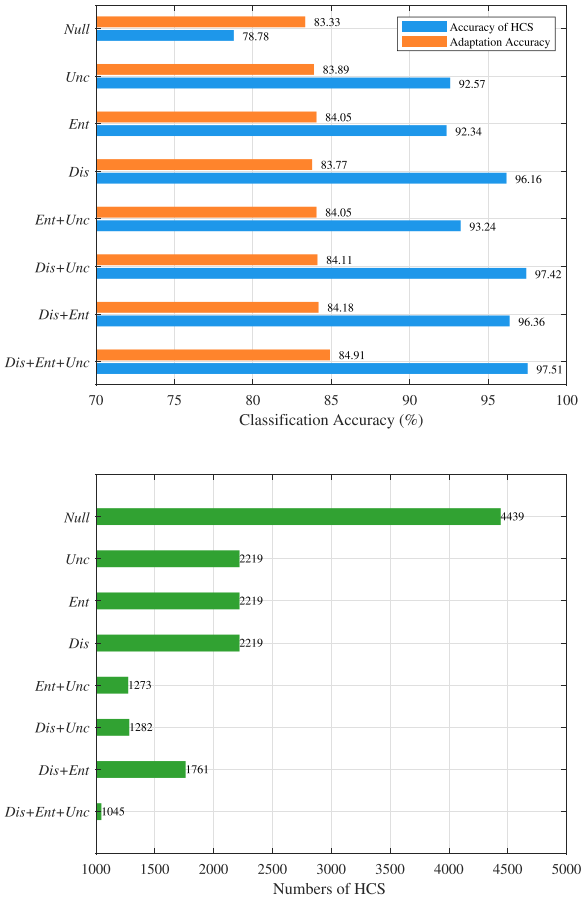


Fig. 8. Effect of HCS selection strategies of Ar→Pr on Office–Home dataset. Up is classification accuracy, and down is the number of HCS.

the necessity of the HCS selection strategy based on evidence theory for target domain adaptation.

Fig. 8 also shows the number of HCSs obtained with different selection schemes. This verifies the effectiveness of each selection criterion. Since the HCS threshold is set as the median overall target domain samples when a single-selection strategy is adopted, the number of HCSs obtained with any one approach is half the total number of target domain data. Therefore, the HCSs selected by the *Unc*, *Ent*, and *Dis* strategies is half the total number of target domain samples. With an increase in selection criteria, the total number of qualified samples will gradually decrease, but the HCS accuracy will increase. The usage of all three selection schemes yields the best results.

5. Conclusions

This paper addresses an SFDA framework for image classification based on evidence theory and a contrastive learning strategy. Based on the inherent properties of feature clustering, we proposed a neighborhood-guided evidence-based contrastive learning scheme that aims to achieve intro-class compactness and inter-class separability. By self-supervised mechanism, the proposed pre-trained source model with transformer layer and evidence analyzer has a better generalization ability in adapting to other domains. The HCS acquisition and joint confidence enhancement mechanisms generate highly deterministic pseudo-labels. In addition, the target domain adaptation is trained dynamically based on the HCS update intermittent training scheme so that the target data has a clear classification decision boundary. Comparative experiments show that UC-SFDA can still achieve the best adaptation effect.

CRediT authorship contribution statement

Dong Chen: Conceptualization, Methodology, Validation, Writing – original draft. **Hongqing Zhu:** Funding acquisition, Project administration, Supervision. **Suyi Yang:** Formal analysis, Validation, Visualization, Writing – review & editing.

Declaration of competing interest

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled.

Data availability

Data will be made available on request.

Acknowledgments

The authors would like to thank the anonymous reviewers and the associate editor for their insightful comments that significantly improved the quality of this paper. This work was supported by the National Natural Science Foundation of China under Grant 61872143.

References

[1] R. Li, Q. Jiao, W. Cao, H.-S. Wong, S. Wu, Model adaptation: Unsupervised domain adaptation without source data, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9641–9650.
[2] J. Liang, D. Hu, J. Feng, Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation, in: International Conference on Machine Learning, 2020, pp. 6028–6039.
[3] S. Tang, Y. Yang, Z. Ma, N. Hendrich, F. Zeng, et al., Nearest neighborhood-based deep clustering for source data-absent unsupervised domain adaptation, 2021, arXiv preprint arXiv:2107.12585.
[4] X. Zhao, Y. Ou, L. Kaplan, F. Chen, J.-H. Cho, Quantifying classification uncertainty using regularized evidential neural networks, 2019, arXiv preprint arXiv:1910.06864.
[5] Y. Du, H. Yang, M. Chen, J. Jiang, H. Luo, C. Wang, Generation, augmentation, and alignment: A pseudo-source domain based method for source-free domain adaptation, 2021, arXiv preprint arXiv:2109.04015.
[6] Z. Qiu, Y. Zhang, H. Lin, S. Niu, Y. Liu, Q. Du, M. Tan, Source-free domain adaptation via avatar prototype generation and adaptation, 2021, arXiv preprint arXiv:2106.15326.
[7] J. Tian, J. Zhang, W. Li, D. Xu, VDM-DA: Virtual domain modeling for source data-free domain adaptation, IEEE Trans. Circuits Syst. Video Technol. 32 (6) (2022) 3749–3760.
[8] G. Yang, H. Tang, Z. Zhong, M. Ding, L. Shao, et al., Transformer-based source-free domain adaptation, 2021, arXiv preprint arXiv:2105.14138.
[9] J. Huang, D. Guan, A. Xiao, S. Lu, Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data, in: Advances in Neural Information Processing Systems, Vol. 34, 2021, pp. 3635–3649.
[10] S. Yang, Y. Wang, J. van de Weijer, L. Herranz, S. Jui, Exploiting the intrinsic neighborhood structure for source-free domain adaptation, in: Advances in Neural Information Processing Systems, Vol. 34, 2021, pp. 29393–29405.
[11] L. Huang, S. Ruan, P. Decazes, T. Denoeux, Lymphoma segmentation from 3D PET-CT images using a deep evidential network, 2022, arXiv preprint arXiv:2201.13078.
[12] B. Yuan, X. Yue, Y. Lv, T. Denoeux, Evidential deep neural networks for uncertain data classification, in: International Conference on Knowledge Science, Engineering and Management, 2020, pp. 427–437.
[13] Z. Tong, P. Xu, T. Denoeux, An evidential classifier based on Dempster-Shafer theory and deep learning, Neurocomputing 450 (2021) 275–293.
[14] A.v.d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, 2018, arXiv preprint arXiv:1807.03748.
[15] S. Dai, Y. Cheng, Y. Zhang, Z. Gan, J. Liu, L. Carin, Contrastively smoothed class alignment for unsupervised domain adaptation, in: Proceedings of the Asian Conference on Computer Vision, 2020.
[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, et al., Attention is all you need, in: Advances in Neural Information Processing Systems, Vol. 30, 2017.
[17] T. Sun, C. Lu, T. Zhang, H. Ling, Safe self-refinement for transformer-based domain adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.
[18] T. Xu, W. Chen, P. Wang, F. Wang, H. Li, R. Jin, Cdtrans: Cross-domain transformer for unsupervised domain adaptation, 2021, arXiv preprint arXiv:2109.06165.
[19] C. Zhang, Q. Zhao, Attention guided for partial domain adaptation, Inform. Sci. 547 (2021) 860–869.
[20] R. Müller, S. Kornblith, G.E. Hinton, When does label smoothing help? in: Advances in Neural Information Processing Systems, Vol. 32, 2019.
[21] M. Sensoy, L. Kaplan, M. Kandemir, Evidential deep learning to quantify classification uncertainty, in: Advances in Neural Information Processing Systems, Vol. 31, 2018.
[22] M. Caron, P. Bojanowski, A. Joulin, M. Douze, Deep clustering for unsupervised learning of visual features, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 132–149.
[23] K. Saenko, B. Kulis, M. Fritz, T. Darrell, Adapting visual category models to new domains, in: Proceedings of the European Conference on Computer Vision, 2010, pp. 213–226.
[24] H. Venkateswara, J. Eusebio, S. Chakraborty, S. Panchanathan, Deep hashing network for unsupervised domain adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 5018–5027.
[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, et al., An image is worth 16 × 16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021.
[26] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, et al., Domain-adversarial training of neural networks, J. Mach. Learn. Res. 17 (1) (2016) 1–35.
[27] K. Saito, K. Watanabe, Y. Ushiku, T. Harada, Maximum classifier discrepancy for unsupervised domain adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 3723–3732.
[28] M. Long, Z. CAO, J. Wang, M.I. Jordan, Conditional adversarial domain adaptation, in: Advances in Neural Information Processing Systems, Vol. 31, 2018.
[29] Y. Zhang, T. Liu, M. Long, M. Jordan, Bridging theory and algorithm for domain adaptation, in: International Conference on Machine Learning, Vol. 97, 2019, pp. 7404–7413.
[30] G. Kang, L. Jiang, Y. Yang, A.G. Hauptmann, Contrastive adaptation network for unsupervised domain adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4893–4902.
[31] Y. Wu, D. Inkpen, A. El-Roby, Dual mixup regularized learning for adversarial domain adaptation, in: Proceedings of the European Conference on Computer Vision, 2020, pp. 540–555.
[32] G. Yang, H. Xia, M. Ding, Z. Ding, Bi-directional generation for unsupervised domain adaptation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 6615–6622.
[33] Y. Jin, X. Wang, M. Long, J. Wang, Minimum class confusion for versatile domain adaptation, in: Proceedings of the European Conference on Computer Vision, 2020, pp. 464–480.
[34] H. Tang, K. Chen, K. Jia, Unsupervised domain adaptation via structurally regularized deep clustering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8725–8735.

[35] S. Cui, S. Wang, J. Zhuo, L. Li, Q. Huang, Q. Tian, Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3941–3950.
[36] R. Xu, P. Liu, L. Wang, C. Chen, J. Wang, Reliable weighted optimal transport for unsupervised domain adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4394–4403.
[37] X. Gu, J. Sun, Z. Xu, Spherical space domain adaptation with robust pseudo-label loss, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9101–9110.
[38] X. Wang, L. Li, W. Ye, M. Long, J. Wang, Transferable attention for domain adaptation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, No. 01, 2019, pp. 5345–5352.
[39] R. Xu, G. Li, J. Yang, L. Lin, Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1426–1435.
[40] Y. Zhang, H. Tang, K. Jia, M. Tan, Domain-symmetric networks for adversarial domain adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5031–5040.
[41] Y. Kim, D. Cho, K. Han, P. Panda, S. Hong, Domain adaptation without source data, 2020, arXiv preprint arXiv:2007.01524.
[42] M. Ishii, M. Sugiyama, Source-free domain adaptation via distributional alignment by matching batch normalization statistics, 2021, arXiv preprint arXiv:2101.10842.
[43] S. Yang, Y. Wang, J. van de Weijer, L. Herranz, S. Jui, Casting a BAIT for offline and online source-free domain adaptation, 2020, arXiv preprint arXiv:2010.12427.
[44] S. Yang, Y. Wang, K. Wang, J. van de Weijer, S. Jui, Local prediction aggregation: A frustratingly easy source-free domain adaptation method, 2022, arXiv preprint arXiv:2205.04183.
[45] N. Ding, Y. Xu, Y. Tang, C. Xu, Y. Wang, D. Tao, Source-free domain adaptation via distribution estimation, 2022, arXiv preprint arXiv:2204.11257.
[46] W. Li, M. Cao, S. Chen, Jacobian norm for unsupervised source-free domain adaptation, 2022, arXiv preprint arXiv:2204.03467.
[47] L. Van Der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) (2008).
[48] A. Chattopadhay, A. Sarkar, P. Howlader, V.N. Balasubramanian, Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks, in: 2018 IEEE Winter Conference on Applications of Computer Vision, 2018, pp. 839–847.