# $E^2$: ENTROPY DISCRIMINATION AND ENERGY OPTIMIZATION FOR SOURCE-FREE UNIVERSAL DOMAIN ADAPTATION

**Anonymous authors**
Paper under double-blind review

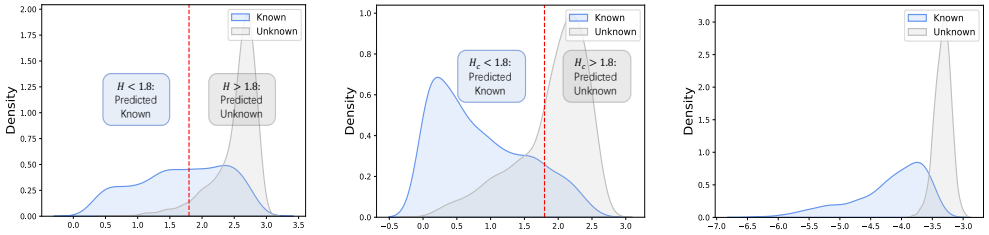
## ABSTRACT

Universal domain adaptation (UniDA) aims to tackle the knowledge transfer problem in the presence of both distribution and category shifts. Most existing UniDA methods are developed based on the accessibility assumption of source-domain data during target model adaptation, which may result in privacy policy violation and source-data transfer inefficiency. To address this issue, we propose a novel source-free UniDA method by confidence-guided entropy discrimination and likelihood-induced energy optimization. The entropy-based separation criterion to determine known- and unknown-class target data may be too conservative for known-class prediction. Thus, we derive the confidence-guided entropy by scaling the normalized prediction score with the known-class confidence, such that much more known-class samples are correctly predicted. Without source-domain data for distribution alignment, we constrain the target-domain marginal distribution by maximizing the known-class likelihood and minimizing the unknown-class one. Since the marginal distribution is difficult to estimate but can be written as a function of free energy, the likelihood-induced loss is changed to an equivalent form based on energy optimization. Theoretically, the proposed method amounts to decreasing and increasing internal energy of known and unknown classes in physics, respectively. Extensive experiments on four publicly available datasets demonstrate the superiority of our method for source-free UniDA.

## 1 INTRODUCTION

Data-driven deep learning models have achieved remarkable success in many computer vision applications such as visual classification (Wang et al., 2020), object detection (Xie et al., 2021) and semantic segmentation (Liu et al., 2021). Since data collected from different sensors or environments may suffer from distribution shift, unsupervised domain adaptation (UDA) (Ganin & Lempitsky, 2015; Li et al., 2021b) is proposed to transfer domain-invariant knowledge from source to target domain without target supervision. Besides distribution shift, the category gap across domains is also an important problem to be addressed in partial domain adaptation (PDA) (Cao et al., 2019), open-set domain adaptation (ODA) (Jing et al., 2021; Liu et al., 2019) and universal domain adaptation (UniDA) (Saito et al., 2020; Saito & Saenko, 2021; You et al., 2019). In PDA (or ODA), the source label set is assumed to be a superset (or subset) of the target label set. Different from PDA and ODA, UniDA is a more practical setting in which there are private label sets in both domains.

UniDA is a challenging problem due to the possible negative transfer caused by the presence of the unknown private classes in each domain. Recently, many research works (Gao et al., 2022; Li et al., 2021a; Saito et al., 2020; Saito & Saenko, 2021; You et al., 2019) have been proposed to solve the challenging but more practical problem of UniDA. These methods learn to separate the shared classes in both domains from the target private classes, such that the distribution shift of the shared classes can be reduced to mitigate the negative transfer. Despite the inspiring progress achieved in UniDA, source-domain data may not be always accessible during the model adaptation stage to train the target model. In practice, the data privacy of the source domain is important and the source-domain dateset may be large-scale constraining the data transmission efficiency. Thus, it becomes necessary to develop source-free UniDA methods without accessing the source-domain data for adapting the pre-trained source model to the target domain.




(a) Entropy (b) Confidence-guided Entropy (c) Free Energy

Figure 1: (a) Entropy, (b) Confidence-guided Entropy and (c) Free Energy distributions of target-domain data w.r.t. the pre-trained source model under the W→A task on the Office-31 dataset. The separation of target known (positive) and unknown (negative) classes can be considered as a binary classification problem. Without labels in the target domain, the optimal separation threshold (i.e. the intersection point) cannot be estimated. Under the same threshold 1.8 (marked in red), the entropy-based recall 0.53 in (a) is significantly lower than 0.88 which is the recall of separation based on the proposed confidence-guided entropy in (b). This means much more known-class samples are correctly predicted as known by using the proposed method, which is beneficial for multi-class classification of known classes. By separating the target data into predicted known- and unknown-class subsets, (c) shows that the known-class energy is lower compared to the unknown-class one. Thus, free energy is used to define the loss function for further improvement.

Under the source-free setting, it is difficult, if not impossible, to align the distributions across domains, as the source-domain data is not available for adversarial training (Ganin et al., 2016) or maximum mean discrepancy (MMD) minimization (Long et al., 2015). Though several source-free domain adaptation methods (Hou & Zheng, 2021; Li et al., 2020; Yang et al., 2021a;b) have been proposed to reduce the distribution discrepancy, they are developed based on the closed-set assumption (i.e. the source label set is equal to the target label set) without considering the category shift. To address the source-free UniDA problem, Kundu et al. (2020b) propose to generate negative samples by image composition of source-domain data which synthesizes the distribution out of the source label set. Then, the target model is adapted from the source model pre-trained with the generated negatives by assigning weights to target samples as similarities to the source domain. Nevertheless, the generated source negatives are still different from the the target private classes which cannot be approximated by the combination of source labels. Hence, data of the (unknown) target private classes may not be well separated from the (known) source-domain classes.

To overcome these limitations, we propose a novel confidence-guided **E**ntropy discrimination and likelihood-induced **E**nergy optimization ($E^2$) method for source-free UniDA. Since entropy measures the classification uncertainty, it is probably higher for unknown and lower for known class. Though the target-domain data could be separated into known- and unknown-class subsets with lower and higher entropy respectively, such separation criterion may be too conservative for known-class prediction as shown in Fig. 1a. In our method, we incorporate the known-class confidence to derive the confidence-guided entropy by scaling the normalized prediction vector with its maximum. Based on the confidence-guided entropy for separation, the number of known-class samples wrongly predicted as unknown can be significantly reduced as illustrated in Fig. 1b. Moreover, we prove the monotonically decreasing property of the confidence-guided entropy w.r.t. the known-class confidence and use it to define the loss function for self-supervised discriminative learning.

Besides confidence-guided entropy discrimination, an innovative method is presented to maximize the known-class likelihood and minimize the unknown-class one based on the concept of free energy in physics. Due to the difficulty in estimating the the marginal distribution, we rewrite the loss function on likelihoods of known and unknown classes to an equivalent form depending on free energy. Since the free energy of in-distribution (known-class) data is lower than that of out-of-distribution (unknown-class) data (Fig. 1c), the loss is designed to decrease and increase the free energy of known-class and unknown-class target data respectively. The overall optimization problem is given by combining confidence-guided entropy discrimination and likelihood-induced energy optimization. Theoretical analysis in physics indicates that internal energy is minimized for known class and maximized for unknown class by optimizing the overall loss function.

The contributions of this work are summarized as follows. 1) We develop a novel source-free UniDA approach based on confidence-guided **E**ntropy discrimination and likelihood-induced **E**nergy optimization ($E^2$). Theoretical analysis shows that minimization of the overall loss function amounts to optimization of internal energy in physics. 2) We derive the confidence-guided entropy to balance the trade-off between classification uncertainty and known-class confidence, such that target data can be better separated into known- and unknown-class subsets. 3) Without source-domain data for distribution alignment, target distribution is constrained by energy optimization which is equivalent to likelihood maximization for known and minimization for unknown class. 4) Extensive experiments on four publicly available domain adaptation datasets demonstrate the effectiveness of our $E^2$ method for source-free UniDA.

## 2 Related Work

**Universal Domain Adaptation (UniDA).** The UniDA setting is first introduced in (You et al., 2019) to address the knowledge transfer problem under both distribution and category shifts. In (Fu et al., 2020), a combination of confidence, entropy and consistency is presented as a uncertainty metric to assign larger weights for more transferable samples during training. The Domain Adaptative Neighborhood Clustering via Entropy optimization (DANCE) (Saito et al., 2020) is proposed to learn the structure of the target domain via self supervision. In (Saito & Saenko, 2021), the One-vs-All Network (OVANet) is trained with the labeled source-domain data and an open-set classifier is adapted to the target domain by minimizing entropy. Different from these works which treat all the target private classes as one unknown category, the Domain Consensus Clustering (DCC) method (Li et al., 2021a) aims at not only identifying the private classes from the shared ones, but also separating the private classes themselves. Nevertheless, these UniDA methods depend on the assumption that source-domain data is accessible during target model adaptation, which may result in privacy policy violation and data transfer inefficiency.

**Source-free Domain Adaptation.** Source-free domain adaptation methods (Kundu et al., 2020a;b; Liang et al., 2020; Yang et al., 2022) have been proposed to address the data privacy and transfer inefficiency issues. Existing source-free domain adaptation methods can be divided into two categories. The first approach (Kundu et al., 2020a;b) aims to improve the generalization ability of the source model by generating negative samples in the hope of approximating the data distribution of target private classes. To reduce the computational overhead in data augmentation, the second approach (Liang et al., 2020) fine-tunes the network parameters for implicit distribution alignment between the target domain and the source model.

**Energy-based models (EBMs).** In EBMs (LeCun et al., 2006), free energy is a quality score assigned to a given input which is lower for observed data and higher for unobserved ones. Recently, EBMs have been employed in a wide range of applications like object recognition (Joseph et al., 2021), out-of-distribution detection (Liu et al., 2020a), generative adversarial network (Zhao et al., 2017) and so on. In (Grathwohl et al., 2020), it has been shown that EBMs can give better performance in terms of improved calibration, out-of-distribution detection, and adversarial robustness. In (Xie et al., 2022; Liu et al., 2020b), free energy is used as a regularization term following the traditional EBMs. Both of these two methods are developed under the setting of closed-set domain adaptation without category shift. They CANNOT be directly employed in the more challenging source-free universal domain adaptation. In this paper, we propose to incorporate free energy for source-free UniDA by decreasing the free energy of the known-class samples and increasing that of the unknown-class ones.

## 3 Method

In universal source-free domain adaptation, the source model $f_0$ is first pre-trained with a labeled source domain $\mathcal{D}_s = \{\mathbf{x}_s, y_s\}$ by minimizing the cross-entropy as in (Liang et al., 2020). In the model adaptation stage, the source model $f_0$ and an unlabeled target domain $\mathcal{D}_t = \{\mathbf{x}_t\}$ are used for training the target model $f$ *without* accessing the source-domain data. Denote the label sets of source and target domain as $\mathcal{C}_s$ and $\mathcal{C}_t$, respectively. The shared label set is $\mathcal{C} = \mathcal{C}_s \cap \mathcal{C}_t$. The source private and target private label sets are the relative complements of $\mathcal{C}$ in $\mathcal{C}_s$ and $\mathcal{C}_t$, respectively, i.e., $\bar{\mathcal{C}}_s = \mathcal{C}_s - \mathcal{C}$ and $\bar{\mathcal{C}}_t = \mathcal{C}_t - \mathcal{C}$. Without label information in the target domain, both the shared label

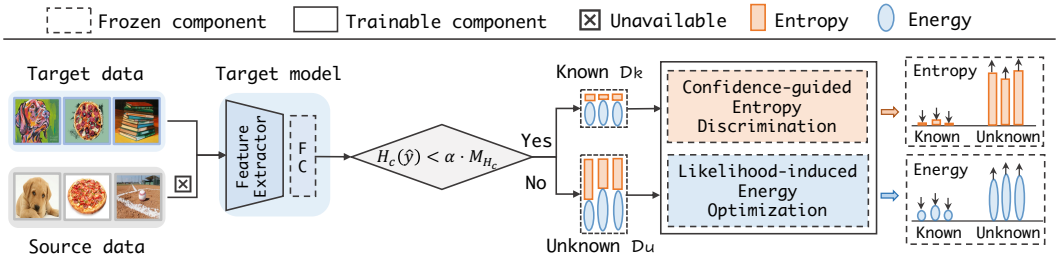



Figure 2: **Overview of the proposed E$^2$ method.** In source-free UniDA, the target model $f$ is initialized with the pre-trained source model. Only the target-domain data is available for training the target model (*without* accessing the source-domain data). The feature extractor is trainable and the fully-connected layers are frozen. We derive the innovative confidence-guided entropy $H_c$ to separate the target-domain data $\mathcal{D}_t$ into two subsets $\mathcal{D}_k$ and $\mathcal{D}_u$ probably containing known- and unknown-class samples. The proposed method includes two components of confidence-guided entropy discrimination and likelihood-induced energy optimization. By training with the two novel components, both the confidence-guided entropy and the energy decrease for known-class and increase for unknown-class samples. The lengths of rectangle and ellipse represent numerical values of the entropy and energy, respectively.

set $\mathcal{C}$ and the private label set $\bar{\mathcal{C}}_t$ are unknown. Denote the number of classes in $\mathcal{C}_s$ as $K = |\mathcal{C}_s|$. The objective is to identify target samples of private classes in $\bar{\mathcal{C}}_t$ as unknown (i.e. the $(|\mathcal{C}_s|+1)$-th class), and classify target samples of shared classes as the corresponding categories in $\mathcal{C}$. The framework of our method is shown in Fig. 2. We propose an innovative E$^2$ approach by combining confidence-guided **E**ntropy discrimination (Sec. 3.1) and likelihood-induced **E**nergy optimization (Sec. 3.2). Theoretical analysis of the proposed method in physics is provided in Sec. 3.3.

### 3.1 Confidence-guided Entropy Discrimination

In our method, we separate the target domain $\mathcal{D}_t$ into two subsets $\mathcal{D}_k$ and $\mathcal{D}_u$ probably containing samples of shared (known) and target private (unknown) classes, respectively. In (Saito et al., 2020), the entropy $H$ of the prediction probability $\hat{y}$ is used for separation of the target-domain data,

$$\mathcal{D}_k = \{\mathbf{x} \in \mathcal{D}_t | H(\hat{y}) < \alpha M_H, \hat{y} = \sigma(f(\mathbf{x}))\}, \mathcal{D}_u = \{\mathbf{x} \in \mathcal{D}_t | H(\hat{y}) > \alpha M_H, \hat{y} = \sigma(f(\mathbf{x}))\} \quad (1)$$

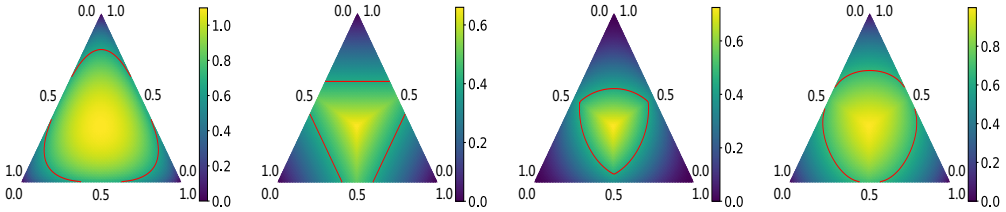
where $\sigma(\cdot)$ denotes the softmax function, $M_H = \log K$ is the maximum value of the entropy and $\alpha \in (0, 1)$ is a hyperparameter of the threshold percentage. Nevertheless, the entropy $H$ may not be discriminative enough for recognizing known and unknown classes in the target domain. For example, if we have two target samples with estimated distributions $(0.5, 0.5, 0)$ and $(0.7, 0.2, 0.1)$, the entropy of $(0.5, 0.5, 0)$ is 0.69 smaller than 0.80 which is the entropy of $(0.7, 0.2, 0.1)$. Entropy measures the classification uncertainty which is larger for unknown class and smaller for known classes. Based on the entropy separation criterion, $(0.7, 0.2, 0.1)$ would not be classified to known classes, if $(0.5, 0.5, 0)$ is assigned to unknown. However, when the known-class confidence $c = \max_i \hat{y}_i$ is used to separate the known and unknown classes, $(0.7, 0.2, 0.1)$ with higher confidence is more likely to belong to known classes, which is not consistent with the entropy-based separation.

A straightforward way to combine the entropy $H$ and the known-class confidence $c$ is the multiplication of $H$ and $1-c$, since both $H$ and $1-c$ are probably lower for known classes and higher for unknown class. The property of using $(1-c)H$ as the separation score is given by Proposition 1.

**Proposition 1.** *Denote $\mathcal{D}_k^H$, $\mathcal{D}_k^{1-c}$ and $\mathcal{D}_k^{(1-c)H}$ as the known-class subsets w.r.t. $H$, $1-c$ and $(1-c)H$, under the same hyperparameter of the threshold percentage $\alpha$. Then, we have,*

$$\mathcal{D}_k^H \subset \mathcal{D}_k^{(1-c)H}, \mathcal{D}_k^{1-c} \subset \mathcal{D}_k^{(1-c)H} \quad (2)$$

The proof of Proposition 1 is provided in Appendix Sec. A.1. According to Proposition 1, if $(1-c)H$ is used as the separation score, more samples will be predicted as known classes compared to the separation based on $H$ or $1-c$, which is also illustrated in Fig. 3. The inclusion relation $\mathcal{D}_k^H \subset \mathcal{D}_k^{(1-c)H}$ means that $\mathcal{D}_k^{(1-c)H}$ contains more samples with higher classification uncertainty



(a) $H$ (b) $1-c$ (c) $(1-c)H$ (d) $H_c$

Figure 3: Distributions of three variables for (a) entropy $H$, (b) $1-c$, (c) $(1-c)H$ and (d) the proposed **confidence-guided entropy** $H_c$, where $c$ is the known-class confidence given by $c = \max_i \hat{y}_i$. The contour of each distribution with the separation threshold equal to 0.6 multiplying the maximum is marked in red. Samples outside and inside the contour are classified as known and unknown, respectively. The area ratios between the estimated known- and unknown-class samples are 0.21 for $H$, 0.92 for $1-c$, 2.02 for $(1-c)H$ and 0.41 for $H_c$. This means entropy $H$ is the most conservative for known-class classification, while a sample is more likely to be predicted as known classes by using $1-c$ or $(1-c)H$ as the separation score. The proposed confidence-guided entropy $H_c$ balances the known- to unknown-class ratios w.r.t. $H$ and $1-c$.

measured by $H$, while $\mathcal{D}_k^{1-c} \subset \mathcal{D}_k^{(1-c)H}$ implies more samples with lower known-class confidence $c$ (higher $1-c$) are in $\mathcal{D}_k^{(1-c)H}$. Since unknown-class samples are probably with higher $H$ or lower $c$, the proportion of true known-class samples in $\mathcal{D}_k^{(1-c)H}$ decreases compared to $\mathcal{D}_k^{H}$ and $\mathcal{D}_k^{1-c}$.

To increase the separation discriminability and balance the trade-off between $H$ and $1-c$, we define the confidence-guided entropy $H_c$ by scaling each $\hat{y}_i$ with $1-c$, i.e.,

$$H_c(\hat{y}) = -\sum_i (1-c)\hat{y}_i \log(1-c)\hat{y}_i \tag{3}$$

The target known- and unknown-class subsets in our method are obtained based on $H_c$ as follows,

$$\mathcal{D}_k^{H_c} = \{\mathbf{x} \in \mathcal{D}_t | H_c(\sigma(f(\mathbf{x}))) < \alpha M_{H_c}\}, \mathcal{D}_u^{H_c} = \{\mathbf{x} \in \mathcal{D}_t | H_c(\sigma(f(\mathbf{x}))) > \alpha M_{H_c}\} \tag{4}$$

where $M_{H_c}$ is the maximum of $H_c$. The sensitivity analysis on hyperparameter $\alpha$ is provided in Appendix Sec. B.1. As shown in Fig. 3, if the proposed $H_c$ is used for separation, the ratio between the estimated known- and unknown-class samples is larger than that based on entropy $H$ but smaller than that based on $1-c$. This ensures that the proposed separation criterion is less conservative than $H$ to recognize known-class data for more discriminative classification. At the same time, the true unknown-class proportion in $\mathcal{D}_k^{H_c}$ would not be too large, since less samples are classified as known compared to the known-class subset $\mathcal{D}_k^{(1-c)H}$ based on $1-c$.

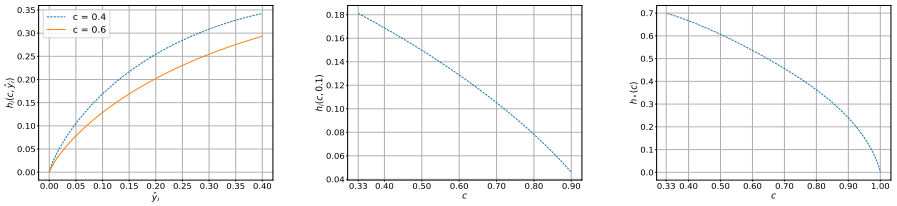
The monotonically decreasing property between the confidence-guided entropy $H_c$ and the known-class confidence $c$ is given in the following proposition:

**Proposition 2.** *The confidence-guided entropy can be rewritten as $H_c(\hat{y}) = \sum_i h(c, \hat{y}_i)$, where $h(c, \hat{y}_i) = -(1-c)\hat{y}_i \log(1-c)\hat{y}_i$. Denote $i_*$ as the class index with the highest prediction probability, i.e., $i_* = \arg\max_i \hat{y}_i$. For $i \neq i_*$, we have $0 \leq \hat{y}_i \leq \min\{c, 1-c\} \leq 0.5$ and the ranges of $c$ is constrained by $\hat{y}_i$, i.e., $\max\{\frac{1}{K}, \hat{y}_i\} \leq c \leq 1-\hat{y}_i$. With a fixed $\hat{y}_i$, $h(c, \hat{y}_i)$ is monotonically decreasing w.r.t. $c$ when $i \neq i_*$. On the other hand, the upper bound of the confidence-guided entropy is a function $h_*(c)$ of the known-class confidence $c = y_{i_*}$, i.e.,*

$$H_c(\hat{y}) \leq h_*(c) = (1-c)^2 \log(K-1) - 2(1-c)^2 \log(1-c) - (1-c)c\log(1-c)c \tag{5}$$

*Equality holds if and only if $\hat{y}_i = \frac{1-c}{K-1}$ for $i \neq i_*$. The upper bound $h_*(c)$ is monotonically decreasing in the range $[\frac{1}{K}, 1]$ of all possible $c$. Thus, the maximum of $H_c$ is $M_{H_c} = h_*(\frac{1}{K})$.*

This proposition is proved in Appendix Sec. A.2. Fig. 4 shows the curves of $h(c, \hat{y}_i)$ with fixed $c$ or $\hat{y}_i$ in the specific range for $i \neq i_*$ as well as the upper bound $h_*(c)$ of all possible $c$. Both Proposition 1 and Fig. 4 illustrate that the confidence-guided entropy $H_c$ is larger (smaller) with smaller (larger) known-class confidence $c$. Without loss of generality, we simplify the symbols of the known- and unknown-class subsets $\mathcal{D}_k^{H_c}$ and $\mathcal{D}_u^{H_c}$ by $\mathcal{D}_k$ and $\mathcal{D}_u$, respectively. For the estimated known-class




(a) $h(c, \hat{y}_i)$ w. fixed $c$ (b) $h(c, 0.1)$ (c) $h_*(c)$

Figure 4: (a) $h(0.4, \hat{y}_i)$ and $h(0.6, \hat{y}_i)$ in the range of $0 \leq \hat{y}_i \leq \min\{c, 1-c\} = 0.4$. (b) $h(c, 0.1)$ in the range of $\hat{y}_i$, i.e., $\frac{1}{3} = \max\{\frac{1}{K}, \hat{y}_i\} \leq c \leq 1 - \hat{y}_i = 0.9$ for $K = 3$. (c) The upper bound $h_*(c)$ in the range $[\frac{1}{K}, 1]$ of all possible $c$ when $K = 3$.

(unknown-class) samples in $\mathcal{D}_k$ ($\mathcal{D}_u$), the entropy uncertainty is minimized (maximized) while the known-class confidence is maximized (minimized). Thus, the loss function $\mathcal{L}_{\text{DIS}}$ for self-supervised discriminative learning based on the confidence-guided entropy $H_c$ is defined as,

$$\mathcal{L}_{\text{DIS}} = \mathbb{E}_{\mathbf{x} \in \mathcal{D}_k} H_c(\hat{y}) - \mathbb{E}_{\mathbf{x} \in \mathcal{D}_u} H_c(\hat{y}) \tag{6}$$

where $\mathbb{E}$ is the expectation operation.

### 3.2 Likelihood-induced Energy Optimization

Without source domain data for distribution alignment, we present an innovative method for known-class likelihood maximization and unknown-class likelihood minimization based on free energy. Denote the known-class marginal distribution as $p$. With the estimated subsets $\mathcal{D}_k$ and $\mathcal{D}_u$, the expected log-likelihood functions are computed by $\mathbb{E}_{\mathbf{x} \in \mathcal{D}_k} \log p(\mathbf{x})$ and $\mathbb{E}_{\mathbf{x} \in \mathcal{D}_u} \log p(\mathbf{x})$ for known and unknown classes, respectively. Since known-class samples are more probable than the unknown under the known-class marginal distribution $p$, the loss function can be defined by maximizing the expected log-likelihood of known class and minimizing the one of unknown class, i.e.,

$$\mathcal{L}_{\text{LL}} = -\mathbb{E}_{\mathbf{x} \in \mathcal{D}_k} \log p(\mathbf{x}) + \mathbb{E}_{\mathbf{x} \in \mathcal{D}_u} \log p(\mathbf{x}) \tag{7}$$

Nevertheless, it is extremely difficult to estimate the marginal distribution $p$ accurately, so that we cannot directly minimize the loss function $\mathcal{L}_{\text{LL}}$ derived based on the expected log-likelihoods.

Instead of estimating the marginal distribution $p$ for optimization, we propose to rewrite $\mathcal{L}_{\text{LL}}$ in eq. (7) by using the physics concept of free energy. For a target sample $\mathbf{x}$ in $\mathcal{D}_k$ or $\mathcal{D}_u$, denote the free energy and the energy w.r.t. the $i$-th class (state) $\omega_i$ as $E(\mathbf{x})$ and $E(\mathbf{x}, \omega_i)$, respectively. As in the energy-based model (Liu et al., 2020a), the relation between the energy functions and the classification model $f$ is derived by connecting the softmax function and the Gibbs distribution. The energy $E(\mathbf{x}, \omega_i)$ of the state $\omega_i$ is equal to the negative of the classification score $f_i(\mathbf{x})$ of the class $\omega_i$, i.e., $E(\mathbf{x}, \omega_i) = -f_i(\mathbf{x})$ (as shown in Appendix Sec. A.3). The free energy[1] quantifying the capacity of a thermodynamic system to do work in physics (Levine, 1978) is given by,

$$E(\mathbf{x}) = -\log \sum_{i=1}^{K} e^{f_i(\mathbf{x})} \tag{8}$$

It has been shown in (LeCun et al., 2006) that the probability distribution $p$ can be written as a function of the free energy $E$, i.e.,

$$p(\mathbf{x}) = \frac{e^{-E(\mathbf{x})}}{\mathcal{Z}} \tag{9}$$

where $\mathcal{Z} = \int e^{-E(\mathbf{x})} d\mathbf{x}$ is an intractable constant for normalization. Substituting eq. (9) into eq. (7) and canceling out the constant $\mathcal{Z}$ which does not depend on $\mathbf{x}$, the loss function $\mathcal{L}_{\text{LL}}$ is changed to $\mathcal{L}_{\text{ELL}}$ based on free energy as follows,

$$\mathcal{L}_{\text{ELL}} = \mathbb{E}_{\mathbf{x} \in \mathcal{D}_k} E(\mathbf{x}) - \mathbb{E}_{\mathbf{x} \in \mathcal{D}_u} E(\mathbf{x}) \tag{10}$$

[1] In contrast, entropy can be considered as the energy in a thermodynamic system that cannot do work.

By minimizing $\mathcal{L}_{\text{ELL}}$ in eq. (10), the energy of known-class samples is minimized for higher known-class marginal probability and vice versa for unknown-class. Denote the partition function of the free energy defined in eq. (8) as $Z(\mathbf{x}) = \sum_{i=1}^{K} e^{f_i(\mathbf{x})}$. The loss function $\mathcal{L}_{\text{ELL}}$ can be rewritten as,

$$\mathcal{L}_{\text{ELL}} = -\log \prod_{\mathbf{x}\in\mathcal{D}_k} Z(\mathbf{x})^{\frac{1}{n_k}} + \log \prod_{\mathbf{x}\in\mathcal{D}_u} Z(\mathbf{x})^{\frac{1}{n_u}} = \log \frac{M_g(\mathcal{D}_u)}{M_g(\mathcal{D}_k)} \tag{11}$$

where $M_g(\mathcal{D}) = \prod_{\mathbf{x}\in\mathcal{D}} Z(\mathbf{x})^{\frac{1}{n}}$ is the geometric mean for $n = n_k$ or $n_u$, $\mathcal{D} = \mathcal{D}_u$ or $\mathcal{D}_k$, $n_k$ and $n_u$ are the numbers of samples in $\mathcal{D}_k$ and $\mathcal{D}_u$, respectively. According to eq. (11), when $\mathcal{L}_{\text{ELL}}$ is minimized, the geometric mean of the partition functions in the estimated known-class subset $\mathcal{D}_k$ is maximized while the one in $\mathcal{D}_u$ is minimized. The overall loss function $\mathcal{L}$ is designed by combining confidence-guided entropy discrimination and likelihood-induced energy optimization, i.e.,

$$\mathcal{L} = \mathcal{L}_{\text{DIS}} + \mathcal{L}_{\text{ELL}} \tag{12}$$

### 3.3 Theoretical Analysis in Physics

Given an input $\mathbf{x}$ and a classification model $f$, the classification score of the class $\omega_i$ is denoted as $z_i = f_i(\mathbf{x})$. The data flow to obtain the prediction probability $\hat{y}_i$ of the class $\omega_i$ from $\mathbf{x}$ is given by,

$$\mathbf{x} \xrightarrow{f} z_i = f_i(\mathbf{x}) = -E(\mathbf{x}, \omega_i) \xrightarrow{\sigma} \hat{y}_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{13}$$

The sum of the entropy $H(\mathbf{x})$ and the free energy $E(\mathbf{x})$ is computed in the following equation,

$$\begin{aligned} H(\mathbf{x}) + E(\mathbf{x}) &= -\sum_{i=1}^{K} \hat{y}_i \log \hat{y}_i - \log \sum_{i=1}^{K} e^{z_i} = -\sum_{i=1}^{K} \hat{y}_i \log \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} - \log \sum_{i=1}^{K} e^{z_i} \\ &= -\sum_{i=1}^{K} \hat{y}_i z_i + \sum_{i=1}^{K} \hat{y}_i \log \sum_{j=1}^{K} e^{z_j} - \log \sum_{i=1}^{K} e^{z_i} = \sum_{i=1}^{K} \hat{y}_i E(\mathbf{x}, \omega_i) = U(\mathbf{x}) \end{aligned} \tag{14}$$

where the internal energy $U$ is the weighted average energy of all the states w.r.t. the occurrence probability $\hat{y}_i$ of $\omega_i$ (Levine, 1978). This means entropy $H$ plus free energy $E$ equals to internal energy $U$. On the other hand, in thermodynamics, the Helmholtz free energy $F$ is defined as,

$$F = U - TS \tag{15}$$

where $T$ denotes the absolute temperature of the surroundings, and $U, S$ are the internal energy and the entropy of the system, respectively. This equation gives similar result that the summation of entropy $S$ multiplying temperature $T$ and Helmholtz free energy $F$ is equal to internal energy $U$.

Therefore, minimizing the overall loss function $\mathcal{L}$ in eq. (12) is equivalent to minimizing and maximizing the internal energy for known and unknown classes, respectively. This can be interpreted as a physical phenomenon. The target domain can be considered as a chemical compound, e.g., $H_2O$. The known and unknown classes are two states of the same chemical compound $H_2O$, e.g., ice and water, respectively. By reducing the internal energy of ice (known-class), it would not melt into water (unknown-class). On the other hand, water (unknown-class) would not become ice (known-class) by heating it. Consequently, known- and unknown-class samples can be better distinguished.

## 4 Experiments

### 4.1 Experiment Setup

**Datasets.** We compare the proposed $\text{E}^2$ method with the state of the art on four publicly available datasets, i.e., **Office-31** (Saenko et al., 2010), **Office-Home** (Venkateswara et al., 2017), **VisDA-C** (**VisDA**) (Peng et al., 2017) and **DomainNet** (Peng et al., 2019). **Office-31** contains 31 object classes in 3 domains, i.e., Amazon, DSLR and Webcam. **Office-Home** consists of 65 object classes in 4 domains, i.e., Art, Product, Clip Art and Real-World. **VisDA** is a large-scale benchmark for object recognition which is composed of 12 classes. **DomainNet** is also larger-scale including about 0.6 million images and 345 classes. We use 3 domains in this dataset following (Fu et al., 2020).

Table 1: Results (%) with class split ($|\bar{\mathcal{C}}^s|/|\mathcal{C}|/|\bar{\mathcal{C}}^t|$=10/10/11) on Office-31 dataset. SF is short for source-free. The best and the second-best results are marked in **bold** and underline, respectively.

| Method | SF | A→D | | A→W | | D→A | | D→W | | W→A | | W→D | | Avg. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | OS | H | OS | H | OS | H | OS | H | OS | H | OS | H | OS | H |
| OSBP | × | 72.9 | 51.1 | 66.1 | 50.2 | 47.4 | 49.8 | 73.6 | 55.5 | 60.5 | 50.2 | 85.6 | 57.5 | 67.7 | 52.3 |
| UAN | × | 86.5 | 59.7 | 85.6 | 58.6 | 85.5 | 60.1 | 94.8 | 70.6 | 85.1 | 60.3 | 98.0 | 71.4 | 89.2 | 63.5 |
| CMU | × | 89.1 | 68.1 | 86.9 | 67.3 | 88.4 | 71.4 | 95.7 | 79.3 | 88.6 | 72.2 | 98.0 | 80.4 | 91.1 | 73.1 |
| ROS | × | - | 71.4 | - | 71.3 | - | 81.0 | - | 94.6 | - | 79.2 | - | 95.3 | - | 82.1 |
| DANCE | × | 91.6 | 78.6 | **92.8** | 71.5 | **92.2** | 79.9 | **97.8** | 91.4 | 91.4 | 72.2 | 97.7 | 87.9 | **93.9** | 80.3 |
| DCC | × | **93.7** | **88.5** | 91.7 | 78.5 | 90.4 | 70.2 | 94.5 | 79.3 | **92.0** | 75.9 | 96.2 | 88.6 | 93.1 | 80.2 |
| OVANet† | × | 84.2 | 84.6 | 74.5 | 78.3 | 67.2 | 76.3 | 96.6 | 95.2 | 77.6 | 82.5 | **99.2** | 95.5 | 83.2 | 85.4 |
| USFDA | ✓ | 88.5 | 85.5 | 85.6 | 79.8 | 87.5 | 83.2 | 95.2 | 90.6 | 86.6 | 81.2 | 97.8 | 88.7 | 90.2 | 84.8 |
| SHOT§ | ✓ | - | 73.5 | - | 67.2 | - | 59.3 | - | 88.3 | - | 77.1 | - | 84.4 | - | 74.9 |
| UMAD | ✓ | - | 79.1 | - | 77.4 | - | 87.4 | - | 90.7 | - | **90.4** | - | **97.2** | - | 87.0 |
| $E^2$ (Ours) | ✓ | 91.6 | 86.8 | 87.2 | **86.3** | 88.9 | **89.7** | 94.9 | **96.0** | 89.3 | 89.6 | 98.5 | 93.4 | 91.7 | **90.3** |

Table 2: HScore (%) with class split ($|\bar{\mathcal{C}}^s|/|\mathcal{C}|/|\bar{\mathcal{C}}^t|$=5/10/50) on Office-Home dataset.

| Method | SF | Ar→Cl | Ar→Pr | Ar→Re | Cl→Ar | Cl→Pr | Cl→Re | Pr→Ar | Pr→Cl | Pr→Re | Re→Ar | Re→Cl | Re→Pr | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| OSBP | × | 39.6 | 45.1 | 46.2 | 45.7 | 45.2 | 46.8 | 45.3 | 40.5 | 45.8 | 45.1 | 41.6 | 46.9 | 44.5 |
| UAN | × | 51.6 | 51.7 | 54.3 | 61.7 | 57.6 | 61.9 | 50.4 | 47.6 | 61.5 | 62.9 | 52.6 | 65.2 | 56.6 |
| CMU | × | 56.0 | 56.9 | 59.2 | 67.0 | 64.3 | 67.8 | 54.7 | 51.1 | 66.4 | 68.2 | 57.9 | 69.7 | 61.6 |
| ROS | × | 60.1 | 69.3 | 76.5 | 58.9 | 65.2 | 68.6 | 60.6 | 56.3 | 74.4 | 68.8 | 60.4 | 75.7 | 66.2 |
| DANCE | × | 61.0 | 60.4 | 64.9 | 65.7 | 58.8 | 61.8 | 73.1 | 61.2 | 66.6 | 67.7 | 62.4 | 63.7 | 63.9 |
| DCC | × | 58.0 | 54.1 | 58.0 | **74.6** | **70.6** | **77.5** | 64.3 | **73.6** | 74.9 | **81.0** | **75.1** | **80.4** | 70.2 |
| OVANet§ | × | 59.7 | **76.9** | 80.0 | 68.8 | 69.1 | 76.2 | 69.6 | 56.9 | **81.0** | 75.5 | 62.0 | 78.6 | 71.2 |
| SHOT§ | ✓ | 32.9 | 29.5 | 39.6 | 56.8 | 30.1 | 41.1 | 54.9 | 35.4 | 42.3 | 58.5 | 33.5 | 33.3 | 40.7 |
| UMAD | ✓ | 61.1 | 76.3 | **82.7** | 70.7 | 67.7 | 75.7 | 64.4 | 55.7 | 76.3 | 73.2 | 60.4 | 77.2 | 70.1 |
| $E^2$ (Ours) | ✓ | **63.2** | 73.9 | 78.8 | 70.1 | 65.6 | 72.9 | **74.9** | 60.8 | **81.0** | 77.1 | 64.2 | 77.3 | **71.7** |

**Evaluation Metric.** We use the average per-class accuracy OS and the HScore (Fu et al., 2020) as the evaluation metric for comparison. The HScore is defined as follows,

$$\text{HScore} = 2 \cdot \frac{\text{OS}^* \cdot \text{UNK}}{\text{OS}^* + \text{UNK}} \tag{16}$$

where $\text{OS}^*$ denotes the average per-class accuracy on known classes and UNK denotes the accuracy on unknown class.

**Implementation Details.** We employ the ResNet50 (He et al., 2016) pre-trained on ImageNet as the backbone and use the PyTorch (Paszke et al., 2019) framework with GeForce RTX 3090 for all the experiments. The source model is pre-trained as in (Liang et al., 2020) with label smoothing. Since the maximum values of the entropy and confidence-guided entropy are almost the same for large $K$ (see proof in Appendix Sec. A.4), we set the separation threshold in eq. (4) as $0.6 \log K$. Stochastic Gradient Descent (SGD) optimizer with momentum $0.9$ is adopted together with the inverse scheduler for learning rate decay. In the stage of source model training, the learning rate is set as $1\text{e}^{-3}$ for Office-31, Office-Home and VisDA datasets. For DomainNet dataset, the learning rate is set as $1\text{e}^{-2}$ for batch normalization and fully-connected layers, and is set as $1\text{e}^{-3}$ for other layers. The batch size 36 is set as default in the source model training stage. During target model adaptation, the training parameters including learning rate $1\text{e}^{-5}$ and batch size 100 are set as default.

**Comparing Methods.** Our method is compared with source-data-accessible and source-free methods. Source-data-accessible methods include OSBP (Saito et al., 2018), UAN (You et al., 2019), CMU (Fu et al., 2020), ROS (Bucci et al., 2020), DANCE (Saito et al., 2020), DCC (Li et al., 2021a) and OVANet (Saito & Saenko, 2021). Source-free methods are USFDA (Kundu et al., 2020b), SHOT (Liang et al., 2020) and UMAD (Liang et al., 2021).

### 4.2 Comparison Results

The OS and HScore results[2] of our method on Office-31 are compared the state of the art in Table 1. Table 2 and Table 3 report the HScore on the Office-Home, DomainNet and VisDA datasets. From

[2]The symbol † means our implementation of the official codes and § refers to results cited from UMAD.

Table 3: HScore (%) on VisDA ($|\bar{\mathcal{C}}^s|/|\mathcal{C}|/|\bar{\mathcal{C}}^t|$=3/6/3) and DomainNet ($|\bar{\mathcal{C}}^s|/|\mathcal{C}|/|\bar{\mathcal{C}}^t|$=50/150/145).

| Method | SF | VisDA | DomainNet | | | | | | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| | | Avg. | P→R | P→S | R→P | R→S | S→P | S→R | |
| OSBP | × | 27.3 | 33.6 | 30.6 | 33.0 | 30.6 | 30.5 | 33.7 | 32.0 |
| UAN | × | 30.5 | 41.9 | 39.1 | 43.6 | 38.7 | 39.0 | 43.7 | 41.0 |
| CMU | × | 34.6 | 50.8 | 45.1 | **52.2** | 45.6 | 44.8 | 51.0 | 48.3 |
| ROS[§] | × | 30.3 | 20.5 | 30.0 | 36.9 | 28.7 | 19.9 | 23.2 | 26.5 |
| DANCE | × | 4.4 | 21.0 | 37.0 | 47.3 | **46.7** | 27.7 | 21.0 | 33.5 |
| DCC | × | 43.0 | 56.9 | 43.7 | 50.3 | 43.3 | 44.9 | 56.2 | 49.2 |
| OVANet[†] | × | 53.1 | 55.5 | **45.8** | 51.6 | 42.4 | 45.6 | **56.6** | **49.6** |
| SHOT[§] | ✓ | 44.0 | 35.0 | 30.8 | 37.2 | 28.3 | 31.9 | 32.2 | 32.6 |
| UMAD | ✓ | 58.3 | **59.0** | 44.3 | 50.1 | 42.1 | 32.0 | 55.3 | 47.1 |
| $E^2$ (Ours) | ✓ | **63.4** | 56.3 | 44.1 | 48.7 | 39.4 | **46.1** | 55.2 | 48.3 |

Table 4: Ablation study on loss functions.

| Ablations | | | Office | | | VisDA | | | DomainNet | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\mathcal{L}_{\text{ENT}}$ | $\mathcal{L}_{\text{DIS}}$ | $\mathcal{L}_{\text{ELL}}$ | OS | OS* | H | OS | OS* | H | OS | OS* | H |
| ✓ | | | 84.2 | 83.8 | 85.2 | 48.6 | 41.8 | 57.0 | 29.8 | 29.4 | 44.0 |
| | ✓ | | 84.8 | 84.5 | 86.2 | 53.0 | 48.8 | 60.0 | 33.0 | 32.7 | 46.9 |
| | ✓ | ✓ | 89.3 ↑5.1 | 89.2 ↑5.4 | 89.6 ↑4.4 | 60.7 ↑12.1 | 59.4 ↑17.6 | 63.4 ↑6.4 | 35.4 ↑5.6 | 35.1 ↑5.7 | 48.7 ↑4.7 |

these results, we can see that the proposed $E^2$ achieves the highest average HScore on the Office-31, Office-Home and VisDA, compared with all the other methods even including the ones accessing the source-domain data for model adaptation. On the Office-31 dataset, the average OS and HScore of our method are $8.5\%$ and $4.9\%$ higher than those of the OVANet which is the source-data-accessible method with the highest HScore. Compared with the competitive source-free universal model adaptation method UMAD, the proposed $E^2$ improves the HScore performance by $3.3\%$ in average on Office-31. Results of the Office-Home dataset show that the proposed method outperforms the state of the art with the highest average HScore. On the VisDA dataset, our method achieves the best performance and is better than UMAD by $5.1\%$ average HScore and significantly outperform other methods (e.g., $10.3\%$ higher than OVANet) no matter source data is available or not. On the DomainNet dataset, our method achieves competitive results compared with all the existing methods and improves over the source-free domain adaptation methods by $1.2\%$ in average HScore.

### 4.3 Ablation Study

Ablation study on different combinations of the loss functions including the entropy loss $\mathcal{L}_{\text{ENT}}$ (by substituting the entropy $H$ for $H_c$ in eq. (6)), the proposed confidence-guided entropy loss $\mathcal{L}_{\text{DIS}}$ and likelihood-induced energy loss $\mathcal{L}_{\text{ELL}}$. Experimental results on the Office-31 (W→A), VisDA and DomainNet (R→P) are shown in Table 4. The baseline results obtained by using the entropy loss $\mathcal{L}_{\text{ENT}}$ for both the target known- and unknown-class samples are recorded in the first row. Results in the second and the third rows show that the performance can be improved by replacing $\mathcal{L}_{\text{ENT}}$ with the proposed confidence-guided entropy loss $\mathcal{L}_{\text{DIS}}$ and adding the energy loss $\mathcal{L}_{\text{ELL}}$. By combining $\mathcal{L}_{\text{DIS}}$ and $\mathcal{L}_{\text{ELL}}$ in the proposed $E^2$, the highest results (OS, OS$^*$ and HScore) are obtained.

## 5 Conclusion

In this paper, we propose a novel source-free UniDA method based on two innovative components of confidence-guided **E**ntropy discrimination and likelihood-induced **E**nergy optimization ($E^2$). The confidence-guided entropy is defined by scaling the normalized prediction score with the known-class confidence, so as to correctly recognize more known-class data. To constrain the target distribution, likelihood is maximized for the known class and minimized for the unknown class by the proposed free energy optimization technique. In-depth theoretical analysis connects our method with the physical concept of internal energy, which uncovers the underlying mechanism on why our method works. Extensive experiments on four domain adaptation datasets demonstrate the superiority of the proposed $E^2$ method compared with the state of the art. Our future work will explore whether the proposed approach is still effective to improve the performance for other challenging problems (e.g., detection, segmentation) under the source-free UniDA setting.

## REFERENCES


Silvia Bucci, Mohammad Reza Loghmani, and Tatiana Tommasi. On the effectiveness of image rotation for open set domain adaptation. In *European Conference on Computer Vision*, pp. 422–438, 2020.

Zhangjie Cao, Kaichao You, Mingsheng Long, Jianmin Wang, and Qiang Yang. Learning to transfer examples for partial domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2985–2994, 2019.

Bo Fu, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Learning to detect open classes for universal domain adaptation. In *European Conference on Computer Vision*, pp. 567–583, 2020.

Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pp. 1180–1189, 2015.

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

Yuan Gao, Peipeng Chen, Yue Gao, Jinpeng Wang, Youngsun Pan, and Andy J. Ma. Hierarchical feature disentangling network for universal domain adaptation. *Pattern Recognition*, 127:108616, 2022.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Yunzhong Hou and Liang Zheng. Visualizing adapted knowledge in domain transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 13824–13833, 2021.

Taotao Jing, Hongfu Liu, and Zhengming Ding. Towards novel target discovery through open-set domain adaptation. In *IEEE International Conference on Computer Vision*, pp. 9302–9311, 2021.

K. J. Joseph, Salman H. Khan, Fahad Shahbaz Khan, and Vineeth N. Balasubramanian. Towards open world object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5830–5840, 2021.

Jogendra Nath Kundu, Naveen Venkat, Ambareesh Revanur, R Venkatesh Babu, et al. Towards inheritable models for open-set domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12376–12385, 2020a.

Jogendra Nath Kundu, Naveen Venkat, Rahul M. V., and R. Venkatesh Babu. Universal source-free domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4543–4552, 2020b.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting Structured Data*, 2006.

Ira Noel Levine. *Physical Chemistry*. MGH, 1978.

Guangrui Li, Guoliang Kang, Yi Zhu, Yunchao Wei, and Yi Yang. Domain consensus clustering for universal domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9757–9766, 2021a.

Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9638–9647, 2020.

Xianfeng Li, Weijie Chen, Di Xie, Shicai Yang, Peng Yuan, Shiliang Pu, and Yueting Zhuang. A free lunch for unsupervised domain adaptive object detection without source data. In *the Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 8474–8481, 2021b.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pp. 6028–6039, 2020.

Jian Liang, Dapeng Hu, Jiashi Feng, and Ran He. UMAD: universal model adaptation under domain and category shift. *https://arxiv.org/abs/2112.08553*, 2021.

Hong Liu, Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Qiang Yang. Separate to adapt: Open set domain adaptation via progressive separation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2927–2936, 2019.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, pp. 21464–21475, 2020a.

Xiaofeng Liu, Bo Hu, Xiongchang Liu, Jun Lu, Jane You, and Lingsheng Kong. Energy-constrained self-training for unsupervised domain adaptation. In *International Conference on Pattern Recognition*, pp. 7515–7520, 2020b.

Yunan Liu, Shanshan Zhang, Yang Li, and Jian Yang. Learning to adapt via latent domains for adaptive semantic segmentation. In *Advances in Neural Information Processing Systems*, pp. 1167–1178, 2021.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pp. 8024–8035, 2019.

Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *IEEE International Conference on Computer Vision*, pp. 1406–1415, 2019.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European Conference on Computer Vision*, pp. 213–226, 2010.

Kuniaki Saito and Kate Saenko. Ovanet: One-vs-all network for universal domain adaptation. In *IEEE International Conference on Computer Vision*, pp. 9000–9009, 2021.

Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *European Conference on Computer Vision*, pp. 153–168, 2018.

Kuniaki Saito, Donghyun Kim, Stan Sclaroff, and Kate Saenko. Universal domain adaptation through self supervision. In *Advances in Neural Information Processing Systems*, pp. 16282–16292, 2020.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

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*, pp. 5018–5027, 2017.

Yulin Wang, Kangchen Lv, Rui Huang, Shiji Song, Le Yang, and Gao Huang. Glance and focus: a dynamic approach to reducing spatial redundancy in image classification. In *Advances in Neural Information Processing Systems*, 2020.

Binhui Xie, Longhui Yuan, Shuang Li, Chi Harold Liu, Xinjing Cheng, and Guoren Wang. Active learning for domain adaptation: An energy-based approach. In *the Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 8708–8716, 2022.

Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *IEEE International Conference on Computer Vision*, pp. 8372–8381, 2021.

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, 2021a.

Baoyao Yang, Andy Jinhua Ma, and Pong C. Yuen. Revealing task-relevant model memorization for source-protected unsupervised domain adaptation. *IEEE Transactions on Information Forensics and Security*, 17:716–731, 2022.

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*, 2021b.

Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Universal domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2720–2729, 2019.

Junbo Jake Zhao, Michaël Mathieu, and Yann LeCun. Energy-based generative adversarial networks. In *International Conference on Learning Representations*, 2017.

# APPENDIX

## CONTENTS

# A SUPPLEMENTAL THEORETICAL ANALYSIS

## A.1 PROOF OF PROPOSITION 1

In this subsection, we present the proof of Proposition 1. The proposition is restated as follows:

**Proposition 1.** *Denote $\mathcal{D}_k^H$, $\mathcal{D}_k^{1-c}$ and $\mathcal{D}_k^{(1-c)H}$ as the known-class subsets w.r.t. $H$, $1-c$ and $(1-c)H$, under the same hyperparameter of the threshold percentage $\alpha$. Then, we have,*

$$\mathcal{D}_k^H \subset \mathcal{D}_k^{(1-c)H}, \mathcal{D}_k^{1-c} \subset \mathcal{D}_k^{(1-c)H} \tag{17}$$

*Proof.* $\mathbf{x} \in \mathcal{D}_k^H$ means $H(\hat{y}) = H(\sigma(f(\mathbf{x}))) < \alpha M_H = \alpha \log K$. To ensure that $\sum_i \hat{y}_i = 1$, $c = \max_i \hat{y}_i \geq 1/K$, which implies $1-1/K$ is the maximum value of $(1-c)$. Thus, $(1-c)H(\hat{y}) \leq (1-1/K)H(\hat{y}) < \alpha(1-1/K)\log K$. Since the maximum value of $(1-c)H(\hat{y})$ is $(1-1/K)\log K$, we have $\mathbf{x} \in \mathcal{D}_k^{(1-c)H}$.

On the other hand, if $\mathbf{x} \in \mathcal{D}_k^{1-c}$, $1-c < \alpha(1-1/K)$. Since the maximum value of $H(\hat{y})$ is $\log K$, we have $(1-c)H(\hat{y}) < \alpha(1-1/K)\log K$. This means $\mathbf{x} \in \mathcal{D}_k^{(1-c)H}$. □

## A.2 PROOF OF PROPOSITION 2

In this subsection, we present the proof of Proposition 2. The proposition is restated as follows:

**Proposition 2.** *The confidence-guided entropy can be rewritten as $H_c(\hat{y}) = \sum_i h(c, \hat{y}_i)$, where $h(c, \hat{y}_i) = -(1-c)\hat{y}_i \log(1-c)\hat{y}_i$. Denote $i_*$ as the class index with the highest prediction probability, i.e., $i_* = \arg\max_i \hat{y}_i$. For $i \neq i_*$, we have $0 \leq \hat{y}_i \leq \min\{c, 1-c\} \leq 0.5$ and the ranges of $c$ is constrained by $\hat{y}_i$, i.e., $\max\{\frac{1}{K}, \hat{y}_i\} \leq c \leq 1-\hat{y}_i$. With a fixed $\hat{y}_i$, $h(c, \hat{y}_i)$ is monotonically decreasing w.r.t. $c$ when $i \neq i_*$. On the other hand, the upper bound of the confidence-guided entropy is a function $h_*(c)$ of the known-class confidence $c = y_{i_*}$, i.e.,*

$$H_c(\hat{y}) \leq h_*(c) = (1-c)^2 \log(K-1) - 2(1-c)^2 \log(1-c) - (1-c)c\log(1-c)c \tag{18}$$

*Equality holds if and only if $\hat{y}_i = \frac{1-c}{K-1}$ for $i \neq i_*$. The upper bound $h_*(c)$ is monotonically decreasing in the range $[\frac{1}{K}, 1]$ of all possible $c$. Thus, the maximum of $H_c$ is $M_{H_c} = h_*(\frac{1}{K})$.*

*Proof.* For $i \neq i_*$, we can prove that $h(c, \hat{y}_i) = -(1-c)\hat{y}_i \log(1-c)\hat{y}_i$ is monotonically decreasing in the range of $\max\{\frac{1}{K}, \hat{y}_i\} \leq c \leq 1-\hat{y}_i$. Since $\hat{y}_i \leq \hat{y}_{i_*}$ and $\hat{y}_i + \hat{y}_{i_*} \leq 1$, we have $0 \leq \hat{y}_i \leq \min\{c, 1-c\} \leq 0.5$ and $c \in [\hat{y}_i, 1-\hat{y}_i]$. On the other hand, $\hat{y}_{i_*} \leq 1 \leq K\hat{y}_{i_*}$, so $c \in [1/K, 1]$ and the range of $c$ is constrained by $\hat{y}_i$ as $\max\{\frac{1}{K}, \hat{y}_i\} \leq c \leq 1-\hat{y}_i$. With a fixed $\hat{y}_i$, the partial derivative of $h(c, \hat{y}_i)$ w.r.t. $c$ is denoted as $h'_c(c, \hat{y}_i)$, and $h'_c(c, \hat{y}_i) = \hat{y}_i \log(1-c)\hat{y}_i + \hat{y}_i$. Then, $h'_c(c, \hat{y}_i)$ is monotonically decreasing and $\forall c \in [\hat{y}_i, 1-\hat{y}_i], h'_c(c, \hat{y}_i) \leq h'_c(\hat{y}_i, \hat{y}_i)$. When $c = \hat{y}_i$, $h'_c(c, \hat{y}_i) = \hat{y}_i \log(1-\hat{y}_i)\hat{y}_i + \hat{y}_i = \hat{y}_i \log e(1-\hat{y}_i)\hat{y}_i$. With $(1-\hat{y}_i)\hat{y}_i \leq 0.25$ and $e(1-\hat{y}_i)\hat{y}_i < 1$, $h'_c(c, \hat{y}_i) < 0$ so $\forall c \in [\hat{y}_i, 1-\hat{y}_i], h'_c(c, \hat{y}_i) < 0$. This means, if $1-\hat{y}_i \geq c_k > c_u \geq \max\{\frac{1}{K}, \hat{y}_i\} \geq \hat{y}_i$, then $h(c_k, \hat{y}_i) < h(c_u, \hat{y}_i)$.

For $i = i_*$, the upper bound of the confidence-guided entropy $H_c$ can be written as a function of $c$ monotonically decreasing when $c \in [1/K, 1]$. Without additional constraint, the range of $c$ is given by $c \in [1/K, 1]$. According to Jensen's inequality, we have,

$$-\sum_{i \neq i_*} (1-c)\hat{y}_i \log(1-c)\hat{y}_i = (1-c)^2 \sum_{i \neq i_*} \frac{\hat{y}_i}{1-c} \log \frac{1}{(1-c)\hat{y}_i} \leq (1-c)^2 \log \frac{K-1}{(1-c)^2} \tag{19}$$

Equality holds if and only if $\hat{y}_i = \frac{1-c}{K-1}$ for $i \neq i_*$. Thus, $H_c(\hat{y}) \leq h_*(c) = (1-c)^2 \log(K-1) - 2(1-c)^2 \log(1-c) - (1-c)c\log(1-c)c$. To prove that $h_*(c)$ is monotonically decreasing with $c \in [1/K, 1]$, we compute the derivative of $h_*(c)$ and obtain,

$$\begin{aligned} h'_*(c) &= -2(1-c)\log(K-1) + 4(1-c)\log(1-c) + 2(1-c) + (2c-1)\log(1-c)c + 2c - 1 \\ &= 2(1-c)\left[-\log(K-1) + 2\log(1-c) - \log(1-c)c\right] + \log(1-c)c + 1 \\ &= 2(1-c)\left[\log(\frac{1}{c} - 1) - \log(K-1)\right] + \log(1-c)ce \end{aligned} \tag{20}$$

For $c \in [1/K, 1]$, $1/c \leq K$, so $\log(1/c-1)-\log(K-1) \leq 0$. On the other hand, $\log(1-c)ce < 0$. As a result, $h'_*(c) < 0$ and $h_*(c_k) < h_*(c_u)$ for $1 \geq c_k > c_u \geq 1/K$. □

### A.3 Relation between Energy $E(\mathbf{x}, \omega_i)$ and Classification Score $f_i(\mathbf{x})$

To ensure that this manuscript is self-contained, we prove the following proposition about the relation between the energy $E(\mathbf{x}, \omega_i)$ and the classification score $f_i(\mathbf{x})$ of the $i$-th class $\omega_i$.

**Proposition 3.** *The energy $E(\mathbf{x}, \omega_i)$ is equal to the negative of the $i$-th classification score $f_i(\mathbf{x})$, i.e., $E(\mathbf{x}, \omega_i) = -f_i(\mathbf{x})$, where $\omega_i$ denotes the $i$-th class.*

*Proof.* According to the definition of Gibbs distribution in statistical mechanics, the posterior $p(\omega_i|\mathbf{x})$ can be given by,

$$p(\omega_i|\mathbf{x}) = \frac{e^{-E(\mathbf{x},\omega_i)}}{\sum_{j=1}^{K} e^{-E(\mathbf{x},\omega_j)}} \tag{21}$$

On the other hand, the posterior $p(\omega_i|\mathbf{x})$ is defined in the softmax operation as,

$$p(\omega_i|\mathbf{x}) = \frac{e^{f_i(\mathbf{x})}}{\sum_{j=1}^{K} e^{f_j(\mathbf{x})}} \tag{22}$$

By comparing eq. (21) and eq. (22), we have $E(\mathbf{x}, \omega_i) = -f_i(\mathbf{x})$. □

### A.4 Relation between Maximums of Entropy $H$ and Confidence-guided Entropy $H_c$

To explain why the upper bound of $H_c$ can be approximated by $\log K$ for the experiments, we prove that maximums of entropy $H$ and confidence-guided entropy $H_c$ are almost the same for large $K$.

**Proposition 4.** *Denote the maximum values of entropy $H$ and confidence-guided entropy $H_c$ as $M_H$ and $M_{H_c}$, respectively. We have $\lim_{K\to+\infty} M_H - M_{H_c} = 0$.*

*Proof.* According to the results in Appendix Sec. A.2, we have $M_{H_c} = h_*(\frac{1}{K})$, where $h_*$ is the upper bound of $H_c$ and $K$ is the number of source classes. By substituting $\frac{1}{K}$ into eq. (18), we have,

$$\begin{aligned}
h_*(\frac{1}{K}) &= (1-\frac{1}{K})^2 \log(K-1) - 2(1-\frac{1}{K})^2 \log(1-\frac{1}{K}) - (1-\frac{1}{K})\frac{1}{K}\log(1-\frac{1}{K})\frac{1}{K} \\
&= (\frac{K-1}{K})^2 \log\frac{K^2}{K-1} - \frac{K-1}{K^2}\log\frac{K-1}{K^2} \\
&= \frac{K-1}{K^2}(\log(\frac{K^2}{K-1})^{(K-1)} - \log\frac{K-1}{K^2}) \\
&= \frac{K-1}{K^2}\log(\frac{K^{(2K-2)}}{(K-1)^{(K-1)}}\frac{K^2}{K-1}) \\
&= \frac{K-1}{K^2}\log\frac{K^{2K}}{(K-1)^K} \\
&= \frac{K-1}{K^2}(2K\log K - K\log(K-1)) \\
&= \frac{2(K-1)}{K}\log K - \frac{K-1}{K}\log(K-1) \\
&= \frac{K-1}{K}\log\frac{K^2}{K-1} \\
&= \log K - \frac{1}{K}\log K + \frac{K-1}{K}\log\frac{K}{K-1}
\end{aligned} \tag{23}$$

On the other hand, the upper bound of entropy $H$ is $\log K$, i.e., $M_H = \log K$. Thus, $\lim_{K\to+\infty} M_H - M_{H_c} = \lim_{K\to+\infty} \frac{1}{K}\log K - \frac{K-1}{K}\log\frac{K}{K-1} = 0$. □

# B ADDITIONAL EXPERIMENTS

## B.1 SENSITIVITY ANALYSIS ON HYPERPARAMETER $\alpha$

For the hyperparameter of the percentage of the separation threshold $\alpha$, we conduct experiments with $\alpha$ in the set of $\{0.5, 0.6, 0.7\}$ on the Office (A→D), Office-Home (Ar→Cl) and VisDA datasets. Results from Fig. 5 show that $\alpha = 0.6$ works well for the three tasks. Thus, the threshold percentage in eq. (4) is set as $0.6$ in all the experiments.

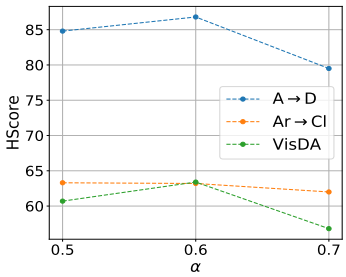
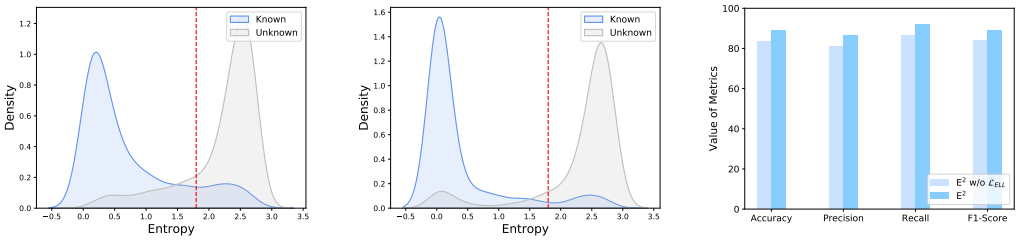



Figure 5: Performance with different $\alpha$.

## B.2 MORE ABLATION RESULTS ON $\mathcal{L}_{\text{ELL}}$

To analyze the effectiveness of the likelihood-induced energy loss $\mathcal{L}_{\text{ELL}}$, we conduct experiments on Office-31 (W→A). Fig. 6a and Fig. 6b show the confidence-guided entropy distributions of target known- and unknown-class data without and with $\mathcal{L}_{\text{ELL}}$ w.r.t the target model. Accuracy, Precision, Recall and F1-Score of the ablation on $\mathcal{L}_{\text{ELL}}$ are presented in Fig. 6c. It can be observed that the binary classification performance to separate the target known- and unknown-class data is improved by adding the proposed likelihood-induced energy loss $\mathcal{L}_{\text{ELL}}$.

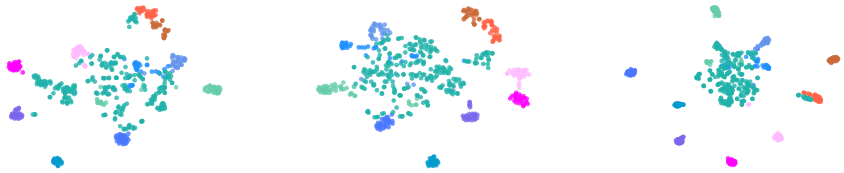



(a) $E^2$ without $\mathcal{L}_{\text{ELL}}$ (b) $E^2$ (c) Quantitative Results

Figure 6: Confidence-guided entropy distributions of the proposed $E^2$ method (a) without $\mathcal{L}_{\text{ELL}}$ and (b) with $\mathcal{L}_{\text{ELL}}$ under the W→A task on Office-31 dataset. (c) Quantitative results of Accuracy, Precision, Recall and F1-Score by using the proposed $E^2$ without or with $\mathcal{L}_{\text{ELL}}$ under the same threshold for binary classification of target known- and unknown-class data.

## B.3 FEATURE VISUALIZATION

To analyze the extracted features visually, Fig. 7 compares the t-SNE (Van der Maaten & Hinton, 2008) results of the source-only model, the OVANet and the proposed $E^2$ method on Office-31 (A→W). From Fig. 7, it can be observed that the target known- and unknown classes are more clearly separated by the proposed $E^2$ compared to the OVANet. Moreover, the clusters of target known classes obtained by the $E^2$ are more compact than those of the OVANet.



(a) Source-only (b) OVANet (c) $E^2$ (Proposed)

Figure 7: t-SNE visualization of the Source-only model, the OVANet and the proposed $E^2$ method. The target unknown-class data is marked in cyan-blue while known-class samples in different colors represent different classes.

### B.4 Comparing Different Strategies for Target Data Separation

To validate the effectiveness of the proposed confidence-guided entropy for target data separation, we compare three different separation strategies based on $(1-c)H$, $(1-c)+H$ and $H_c$, respectively. Experiments are conducted on Office (A→D), Office-Home (Ar→Cl) and VisDA. For $(1-c)+H$, the entropy is normalized by dividing by the range, i.e., $H/\log K$. The threshold is set as $0.6(1-1/K)\log K$ for $(1-c)H$, $0.6(2-1/K)$ for $(1-c)+H$ and $0.6\log K$ for $H_c$. Results from Fig. 8 illustrates that the HScore of the separation strategy based on the proposed $H_c$ is higher than that based on $(1-c)H$ or $(1-c)+H$.

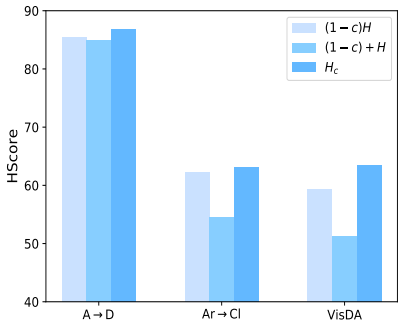



Figure 8: HScore (%) of different strategies for target data separation.

### B.5 Comparison under Different Domain Adaptation Settings

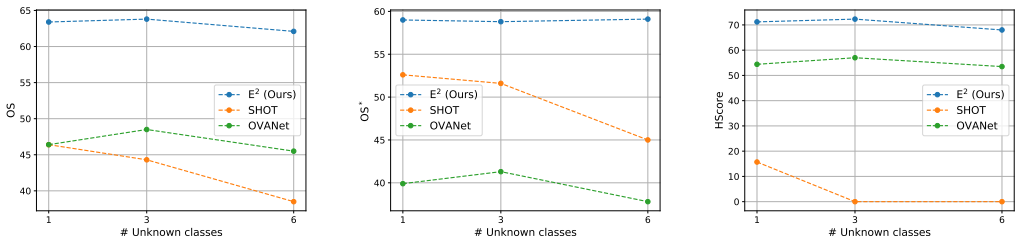
In this experiment, our method is compared with source-accessible and source-free methods under open-set (ODA), partial- and closed-set domain adaptation (PDA and CDA). Source-accessible method is (OVANet) (Saito & Saenko, 2021). Source-free methods are: Source Hypothesis Transfer (SHOT) (Liang et al., 2020), Inheritable models (Inheritune) (Kundu et al., 2020a) and Universal

Table 5: HScore (%) with class split ($|\bar{\mathcal{C}}^s|/|\mathcal{C}|/|\bar{\mathcal{C}}^t|$=0/10/11) on Office-31 dataset under source-free open-set domain adaptation. The best and the second-best results are marked as **bold** and underlined, respectively.

| Method | A→D | A→W | D→A | D→W | W→A | W→D | Avg. |
|---|---|---|---|---|---|---|---|
| SHOT | 77.0 | 70.3 | 66.3 | 83.7 | 58.6 | 84.6 | 73.5 |
| Inheritune | 73.3 | 76.6 | 80.1 | 87.7 | **81.3** | 90.7 | 81.6 |
| USFDA | 80.4 | 69.9 | **81.7** | 92.3 | 78.5 | **93.7** | 82.8 |
| $E^2$ (Ours) | **81.1** | **85.4** | 77.8 | **95.1** | 76.4 | 90.1 | **84.3** |

Table 6: HScore (%) with class split ($|\bar{\mathcal{C}}^s|/|\mathcal{C}|/|\bar{\mathcal{C}}^t|$=0/25/40) on Office-Home dataset under source-free open-set domain adaptation.

| Method | Ar→Cl | Ar→Pr | Ar→Re | Cl→Ar | Cl→Pr | Cl→Re | Pr→Ar | Pr→Cl | Pr→Re | Re→Ar | Re→Cl | Re→Pr | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SHOT | 39.8 | 41.0 | 47.8 | 54.4 | 48.1 | 44.5 | **57.6** | 42.2 | 47.4 | 59.1 | 41.5 | 43.5 | 47.2 |
| Inheritune | 56.1 | 57.5 | 67.2 | 50.6 | 58.4 | 62.7 | 46.8 | 50.7 | 63.2 | 54.5 | 53.5 | 63.9 | 57.1 |
| USFDA | 46.8 | 59.3 | 65.5 | 48.1 | 59.0 | 65.7 | 56.8 | 50.4 | 67.1 | 56.1 | 48.9 | **68.9** | 57.7 |
| E$^2$ (Ours) | **58.0** | **65.3** | **71.7** | **58.7** | **64.1** | **66.3** | 56.8 | **52.1** | **71.5** | **67.0** | **58.5** | 68.6 | **63.2** |



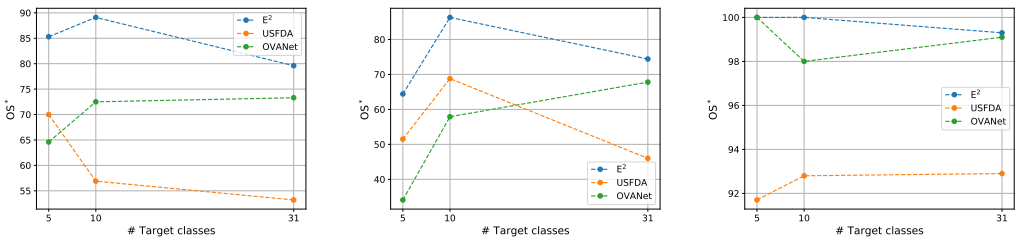

(a) OS (b) OS* (c) HScore

Figure 9: Results (%) on VisDA dataset with different number of private (unknown) classes in the target domain under open-set domain adaptation.

Source-free Domain Adaptation (USFDA) (Kundu et al., 2020b). Results in this section are obtained by the implementation of their official codes.

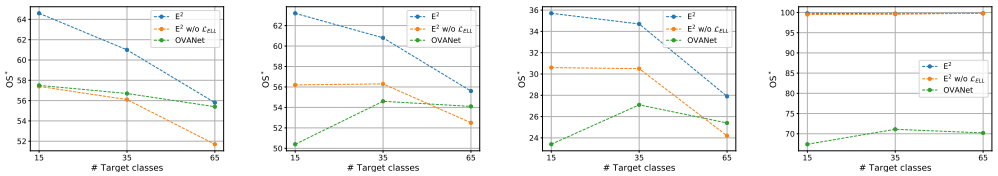
**Open-set Domain Adaptation.** For ODA, results of HScore on the Office-31 and Office-Home datasets are recorded in Table 5 and Table 6, respectively. Results of OS, OS* and HScore on the VisDA dataset are shown in Fig. 9 with varying number of target unknown classes (# Unknown classes=6 is the usual class split case for ODA). From Table 5, it can be observed that the proposed E$^2$ achieves the highest average HScore on Office-31 and Office-Home datasets compared with the state of the art. The average HScore of our method is 2.7% higher than the Inheritune which is tailored-made for open-set source-free domain adaptation. Table 6 shows that the proposed E$^2$ achieves the highest average HScore which is 6.1% higher than the Inheritune. In Fig. 9, the results including OS, OS* and HScore under different openness show that the proposed method gives a notable improvement over SHOT and OVANet.

**Partial- and Closed-Set Domain Adaptation.** For PDA and CDA, the results of OS* on the Office-31 (A→D, A→W and W→D) and Office-Home (Ar→Pr, Cl→Re, Pr→Cl and Re→Pr) datasets under different openness are shown in Fig. 10 and Fig. 11, respectively. When the number of target classes is 31 for Office-31 and 65 for Office-Home dataset, it becomes the CDA problem with the same set of classes across domains. Fig. 10 shows that the proposed method E$^2$ achieves the highest OS* compared with the USFDA and OVANet under different numbers of target classes. In Fig. 11, the proposed method E$^2$ achieves higher OS * than the OVANet. It can be observed that the proposed




(a) A→D (b) A→W (c) W→D

Figure 10: OS* (%) on Office-31 dataset with different number of target classes in the target domain under partial- and closed-set domain adaptation.




(a) Ar→Pr (b) Cl→Re (c) Pr→Cl (d) Re→Pr

Figure 11: OS* (%) on Office-Home dataset with different number of target classes in the target domain under partial- and closed-set domain adaptation.

method $E^2$ without the likelihood-induced energy loss $\mathcal{L}_{\text{ELL}}$, i.e., $E^2$ w/o $\mathcal{L}_{\text{ELL}}$, is even better than the OVANet in some cases under the performance metric of OS* .