Energy-based Self-Training and Normalization for Unsupervised Domain Adaptation

Samitha Herath[1†] Basura Fernando[2] Ehsan Abbasnejad[3] Munawar Hayat[1] Shahram Khadivi[4]
Mehrtash Harandi[1] Hamid Rezatofighi[1] Gholamreza Haffari[1]
[1]Monash University, Australia [2]A*STAR, Singapore [3]The University of Adelaide, Australia [4]eBay Inc.
[†]samitha.herath1@monash.edu

Abstract

*We propose an Unsupervised Domain Adaptation (UDA) method by making use of Energy-Based Learning (EBL) and demonstrate **1.** EBL can be used to improve the instance selection for a self-training task on the unlabelled target domain, and **2.** alignment and normalizing energy scores can learn domain-invariant representations. For the former, we show that an energy-based selection criterion can be used to model instance selections by mimicking the joint distribution between data and predictions in the target domain. As per learning domain invariant representations, we show that stable domain alignment can be achieved by a combined energy alignment and an energy normalization process. We implement our method in consistent with the vision-transformer (ViT) backbone and show that our proposed method can outperform state-of-the-art ViT based UDA methods on diverse benchmarks (DomainNet, Office-Home, and VISDA2017).*

1. Introduction

The recent progress in Unsupervised Domain Adaptation (UDA) methods depend on two key factors: the capacity to learn discriminatively from unlabelled target data and the potential to achieve domain alignment without compromising the integrity of the representation (*i.e.*, avoiding degenerated representations). In this study, we propose a novel approach for UDA that leverages the energy-based interpretation of discriminative classifiers [11, 16, 36]. We show how energy-based learning can be used to generate a better self-training signal when learning discriminatively from the unlabelled target domain data. To this end, we demonstrate that the representation integrity during domain alignment can be maintained with a proposed novel normalization process for the energy-based learning framework. We show that our energy-based self-training, the energy normalization process together with free-energy alignment proposed in [36] forms a seamless energy-based learning framework for UDA on top of the transformer [7] backbone. Our results confirm that this proposed framework outperforms state-of-the-art UDA methods on established UDA benchmarks.

Prior work has emphasized the significance of a discriminative objective in the target domain data for UDA. For example, [28] employs entropy minimization principles [10] as a discriminative loss on the unlabelled data. Prabhu *et al*. [23] illustrate the effectiveness of min-max learning of the entropy loss function when utilized to learn from the unlabelled target domain data. The recent transformer-based technique presented in [24] proposes a self-training task that involves pseudo-labelling and learning augmentation masks from the data. Nonetheless, these approaches only make use of the consistency of the conditional distribution of the labels (*i.e.*, pseudo labels) given the unlabelled data to develop the self-training task. It is evident that the joint distribution between data and labels captures a stronger relationship between them [25]; but an utterly involved choice to be considered. For instance, modelling the joint distribution will require the help of computationally exhaustive generative modelling techniques [37, 38]. In this work, we show how we can mimic the behaviour of the joint distribution modelling to generate an informative self-training signal by using the energy-based learning concepts for classifiers.

With the current success of the transformers [33] in both vision and NLP [6], we focus our efforts on examining the proposed approach using the vision transformer (ViT) [7] backbone. As such, the quality of training depends on the capacity of the self-attention parameters to accurately emphasize related information in the training data. However, for UDA, due to the domain shift between the source and the target data, the self-attention parameters learnt through the source domain supervision may not well align with all target data points. Here, we propose a selection criterion to decide on instances that are compatible with the attention of the transformer model. In our formulation, we use the relationship between free-energy to the marginal density of the data. Thereafter, we show that by a careful combination of such marginal density-based instance selections with

prediction consistency instance selections can lead to better self-training performance.

We make use of the concept of minimizing free-energy bias [36] across the domains to learn domain invariant representations. We integrate free-energy alignment (FEA) of [36] into a module, SCAL (coined from the terms **SC**ore **AL**igment) that can be used in a straightforward way with the classification layer (*i.e.*, logits layer) of a Deep Neural Network (DNN). Moreover, we show that the free-energy based domain alignment can be further stabilized using a normalization process applied on the free-energy and energy scores. We show that this normalization helps to keep the integrity of the features while improving the domain alignment. We coin our novel normalization process into a module, SCON (coined from the terms **SCO**re **N**ormalization) that can be attached to the classifier outputs.

Our choice of distribution alignment, normalization, and joint distribution modeling for self-training is inspired by the principles of energy-based learning. We are the first to combine the concepts of energy-based learning, domain adaptation, self-training, and the ViT backbone. In our experiments, we show that this combination enables us to outperform state-of-the-art unsupervised domain adaptation methods on challenging UDA benchmarks.

- We propose a self-training task that emulates the joint distribution of the data and the predictions using the concepts of free-energy and energy-based learning.
- We propose a novel energy-normalization module, SCON to achieve stable domain alignment when combined with free-energy alignment [36].
- We outperform state-of-the-art methods on established benchmarks (DomainNet [21], OfficeHome [34], VISDA2017 [22]) with relative improvement of $2.2\%-3.3\%$ compared to the best performing method.

2. Background

In this section, we present a brief introduction to our problem setting *i.e.*, UDA, energy-based learning and draw attention to its essential attributes. More specifically, we concentrate on explaining the concept of energy-function, the training objective, and the inference rule within our scope of energy-based learning.

Problem setting. We consider the case where we have access to labelled data points, $(\boldsymbol{x}_j^{(s)}, y_j^{(s)}) \in \mathcal{D}_s$ for $j = 1, 2, 3, \cdots, N_s$ from a source domain and unlabelled data points $\boldsymbol{x}_i^{(t)} \in \mathcal{D}_t$ for $i = 1, 2, 3, \cdots, N_t$ from a target domain. We use the letters "*s*" and "*t*" to refer to **s**ource and **t**arget domains, respectively. As per the UDA protocol, we assume that all data points from the source and target domains share the same set of classes, $\mathcal{C} = \{1, 2, 3, \cdots, c\}$.

Energy-function and inference rule. An energy-based model is built on top of a scalar-valued function $E(\boldsymbol{x}, y)$ called the energy-function that measures the compatibility between an observed variable $\boldsymbol{x} \in \mathbb{R}^{\ell}$[1] and a prediction, $y \in \mathcal{C}$. Formally, $E : \mathbb{R}^{\ell} \times \mathcal{C} \rightarrow \mathbb{R}$. The inference rule for the energy-based model is given by minimizing the energy function over the set of possible predictions $\mathcal{Y}$, i.e., the class labels as in,

$$y^* = \arg\min_{y \in \mathcal{C}} E(\boldsymbol{x}, y). \tag{1}$$

Here, y^* denotes the predicted label. As such, we make the following Remark 1 on the energy-function construction.

Remark 1 *We construct the energy-function, E by taking the negated output of the logits layer (*i.e.*, layer preceding the softmax). It should be noted that this negation operation makes our DNN output consistent with the training and inference criterion we adopt for energy-based learning.*

Free-energy. We use the following definition of free-energy, F in our energy-based learning framework:

$$F(\boldsymbol{x}) = -\log \sum_{c \in \mathcal{Y}} \exp(-\tau E(\boldsymbol{x}, c)). \tag{2}$$

Here, τ is the temperature parameter. Note, the term free-energy is obtained by marginalizing the energy-function outputs across all possible prediction classes. As such, we make the following Remark 2 on the relationship between free-energy and the marginal probability score, $p(\boldsymbol{x})$.

Remark 2 *The probability density of $\boldsymbol{x}$ can be expressed through free-energy, F according to [36]:*

$$p(x) = \frac{\exp(-F(x))}{\sum_{x \in \mathcal{D}_t} \exp(-F(x))}. \tag{3}$$

As such, for two given datapoints $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, the inequality, $F(\boldsymbol{x}_1) < F(\boldsymbol{x}_2)$ implies that the datapoint $\boldsymbol{x}_1$ is occurring from a more dense region compared to $\boldsymbol{x}_2$ w.r.t. the marginal distribution, $p(x)$ parameterized by the energy-function parameters.

Our motivation for using energy-based learning comes in two-fold. Firstly, the capacity of the free-energies to model a better instance selection strategy for self-training. Secondly, the free-energy bias can serve as a disparity measure for domain disparity.

[1] Here, ℓ = height $\times$ width $\times$ #channels of an input image.

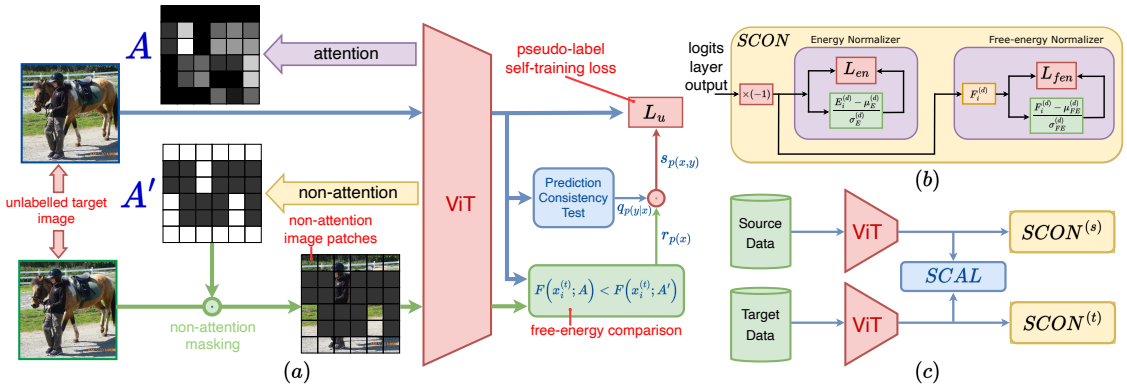

Figure 1. **(a).** The schematic diagram of the proposed free-energy based instance selection criterion. For a given unlabelled target training image, we compute both attention, $\boldsymbol{A}_s$ and non-attention, $\boldsymbol{A}'_s$ regions. To assert the computed attention, $\boldsymbol{A}_s$ is meaningful, we compare the free-energy scores. Our free-energy score based instance selections are combined with consistency based instance selections. **(b).** A schematic of the proposed normalizer module, SCON. SCON consists of two normalization modules, namely for normalizing energy and free-energy scores. **(c).** Integration schematic of SCAL and SCON modules. Note, we use two SCON modules for each domain.

3. Self-training using Energy-based Instance Selection

Recent work of Prabhu *et al*., [24] proposes a self-training task where a target domain instance is decided to be 'reliable' or 'unreliable' depending on the consistency in its predictions. More formally, they consider the quality of the conditional distribution, $P_t(y|\boldsymbol{x};\boldsymbol{A}_s)$. Here, we use the subscript "t" to indicate the target domain and the notation $\boldsymbol{A}_s$ to indicate that this conditional distribution is parameterized by an attention map, $\boldsymbol{A}_s$. We use the subscript "s" to indicate that the attention parameters are trained using a supervised signal from the source domain as in the UDA setting[2]. Despite not being ViT based, the self-training methods discussed in seminal works SENTRY [23], FixMatch [30], UnbiasedTeacher [18] are built on top of a similar motive. In other words, such methods only consider the quality of the conditional distribution of predictions when formulating their instance selection criterion.

To this end, we conjuncture that a stronger selection criterion should be capable of modeling the joint distribution of the prediction and instance, $P_t(\boldsymbol{x}, y; \boldsymbol{A}_s)$. In the following discussion, we explain details of how we benefit from our energy-based learning method to formulate such an instance selection criterion.

3.1. Using Free-energy Estimates for Density based Instance Selection

For models that use self-attention, the quality of the training depends on the capacity of the self-attention parameters to accurately emphasize the relevant information from the input data. However, in UDA, due to the domain mismatch, it is fair to assume that a considerable portion of the target domain instances might not be compatible with the model's self-attention parameters. As such, a self-training signal applied on-top of the model is susceptible to be noisy for such incompatible instances from the target domain.

Having this in mind, we propose to probe the probability score, $p_t(\boldsymbol{x}_i^{(t)};\boldsymbol{A}_s)$[3] given an instance, $\boldsymbol{x}_i^{(t)} \in \mathcal{D}_t$ as an indicator for the compatibility of a target instance with the self-attention parameters learnt on the source domain. In other words, a higher $p_t(\boldsymbol{x}_i^{(t)};\boldsymbol{A}_s)$ score indicates better compatibility between the attention map $\boldsymbol{A}_s$ with the considered target domain datapoint, $\boldsymbol{x}_i^{(t)}$.

To this end, we benefit from the energy-based learning to estimate the marginal probability scores using free-energy (*see* Remark 2). As such, we attempt to build a comparison rule to compare the compatibility of any two given attention maps, $\boldsymbol{A}_s$ and $\boldsymbol{A}'_s$ between the datapoint of interest. For instance, in the case that the attention map $\boldsymbol{A}_s$ has a better compatibility than $\boldsymbol{A}'_s$, it could be said that:

$$p_t(\boldsymbol{x}_i^{(t)};\boldsymbol{A}_s) > p_t(\boldsymbol{x}_i^{(t)};\boldsymbol{A}'_s) \Leftrightarrow F(\boldsymbol{x}_i^{(t)};\boldsymbol{A}_s) < F(\boldsymbol{x}_i^{(t)};\boldsymbol{A}'_s). \tag{4}$$

Although it is straightforward to compute the attention parameter $\boldsymbol{A}_s$, it is not clear to decide on a meaningful set of alternative attention parameters, $\boldsymbol{A}'_s$. In the following section we discuss details of our choices for $\boldsymbol{A}_s$, and $\boldsymbol{A}'_s$.

3.1.1 Computing Attention

We estimate $\boldsymbol{A}_s$ by averaging the self-attention maps across all the attention heads of the final transformer block of our

[2]In reality the attention in ViT models are also dependent on the input, $\boldsymbol{x}$. In other words, the attention is a function, $\boldsymbol{A}_s(\boldsymbol{x})$. However, for clarity, we consider, $\boldsymbol{A}_s$ to depend only on the ViT model's parameters.

[3]We use the notation p_t to distinguish the probability score against the distribution P_t.

ViT model. More specifically, for the final transformer block of our model, we compute the attention parameters, $\boldsymbol{A}_s$, by considering the similarity of the class token with each patch-token as per:

$$\boldsymbol{A}_s = \frac{1}{\omega}\sum_{i=1}^{\omega} \text{softmax}\big(q_{[\text{cls}]}K^T/\sqrt{D}_\omega\big). \tag{5}$$

Here, $K \in \mathbb{R}^{N\times D_\omega}$ and class token query, $q_{[cls]} \in \mathbb{R}^{D_\omega}$ with N being the number of patches of the input target domain image, $\boldsymbol{x}_i^{(t)}$. The dimensionality of the attention-head embeddings is given by D_ω. We compute the set of final attention parameters by averaging across ω attention heads.

Thereby, we speculate that the condition in (4) will be satisfied for a meaningful self-attention $\boldsymbol{A}_s$, provided that $\boldsymbol{A}'_s$ indicates *non-attentions*. In other words, here we consider the case where $\boldsymbol{A}'_s$ is computed to find non-attention regions. We formulate our non-attention regions as per,

$$\boldsymbol{A}'_s = \frac{1}{\omega}\sum_{i=1}^{\omega} \text{softmin}\big(q_{[\text{cls}]}K^T/\sqrt{D}_\omega\big). \tag{6}$$

Note, here we use *softmin* to emphersize the dissimilarity between the $q_{[cls]}$ and patch-token key values, $\boldsymbol{K}$. In Fig. 1(a) we provide a schematic of how attention and non-attention patches are used for instance selection.

3.2. Combining Density based Instance Selection with Conditional Selections

In section 3.1 we established our criterion for instance selection using the free-energy. As such, for a given set of target domain instances, $\{\boldsymbol{x}_i^{(t)}\}_{i=1}^{N_t}$ we compute the instance selections tuple, $\mathcal{R}_{p(x)} = (r_1, r_2, \cdots, r_{N_t})$ with,

$$r_i = \mathbb{I}\big[F(\boldsymbol{x}_i^{(t)}; \boldsymbol{A}_s) < F(\boldsymbol{x}_i^{(t)}; \boldsymbol{A}'_s)\big]. \tag{7}$$

Here, $\mathbb{I}$ is the indicator function, which returns 1 if the free-energy inequality is satisfied, and 0 otherwise. The intuition of (7) is to select instances where attention, $\boldsymbol{A}_s$ is more meaningful than the attention, $\boldsymbol{A}'_s$. *i.e.*, instances from dense data regions where the model has learned meaningful attentions. We use the subscript $p(x)$ in $\mathcal{R}_{p(x)}$ to reflect that these instance selections consider the marginal distribution density scores. Hereby, to mimic the behavior of the joint distribution, we incorporate instance selections based on an existing prediction consistency-based method. More formally, in our implementations, we use the prediction consistency-based selections of PACMAC [24] to build a tuple, $\mathcal{Q}_{p(y|x)} = (q_1, q_2, \cdots, q_{N_t})$. Here q_i will take the value 1 for an instance with a consistent label prediction under PACMAC's criterion. Thereby, to model the joint distribution based selection, we apply an element wise product rule to obtain our final instance selections:

$$\mathcal{S}_{p(x,y)} = \mathcal{R}_{p(x)} \odot \mathcal{Q}_{p(y|x)}. \tag{8}$$

Finally, we restrict the instances to be considered for self-training only to those indicated by the tuple $\mathcal{S}_{p(x,y)} = (s_1, s_2, \cdots, s_{N_t})$. To be precise, we construct our pseudo-labeled self-training loss, $\mathcal{L}_u$ as follows:

$$\mathcal{L}_u = \frac{1}{|\mathcal{S}^+_{p(x,y)}|}\sum_{i=1}^{N_t} s_i \times h_{ce}(\boldsymbol{x}_i^{(t)}, \hat{y}_i). \tag{9}$$

Here, we use, h_{ce} to represent the softmax cross-entropy loss computed using a target domain instance, $\boldsymbol{x}_i^{(t)}$ and its pseudo-label, $\hat{y}_i$. The set $\mathcal{S}^+_{p(x,y)}$ is defined as $\{s \in \mathcal{S}_{p(x,y)} : s = 1\}$ to contain the selected instances for self-training (*see* Fig. 1(a) for the instance selection schematic.).

4. Learning Domain Invariant Features using Energy Alignment and Normalization

In the following sections, we present our approach to learning domain invariant representations by aligning and normalizing energy scores. Specifically, we employ two separate modules: SCAL, based on the free-energy alignment method of [36], and SCON, a novel normalization module.

4.1. SCAL: SCore ALignment for Learning Domain Invariant Representations

We built on-top of prior work [11, 36] that shows the free-energy score differences between the source and target domain data can be used to represent the covariate shift between the two domains. This shift is termed as the free-energy bias. More interestingly, it has been shown that the minimization free-energy bias leads to domain invariant representations. Formally, free-energy alignment (FEA) [36] has been proposed to reduce this bias in free-energy. We realize free-energy alignment by minimizing the following margin-loss proposed in [36]:

$$\mathcal{L}_{ea} = \sum_{i=1}^{N_t} \max(0, F(\boldsymbol{x}_i^{(t)}) - \tilde{\mu}_{FE}). \tag{10}$$

Here, $\tilde{\mu}_{FE}$ is a moving average of source domain free-energy scores computed for each mini-batch.

However, even though free-energy alignment promises domain-invariant properties, we observe that using FEA alone is a demanding task. Free-energy scores are unbounded in nature and aligning unbounded scores often leads to unstable training. Therefore, we propose a novel normalizer module, SCON to make FEA tractable.

4.2. SCON: SCOre Normalization for Stable Free-energy Alignment

In order to meet the inference criterion for energy-based learning, we define the energy-function as the negated out-

puts at the logits layer (*see* Remark 1). This results in unbounded outputs, or energy-scores. Additionally, the free-energy scores are also unbounded due to their construction. However, training SCAL (*see* section 4.1) to match such unbounded score distributions can be a challenging learning objective to achieve.

Normalization operations such as LRN [15], BatchNorm [13], InstanceNorm [32], Layer-Norm [2] have been proposed in DNN literature to manage such situations effectively. Similarly, we conjuncture that the quality of the free-energy alignment can be improved by a statistical normalization. We coin this normalization process as SCON to attribute that it is a **SCO**re **N**ormalization applied on energy-function outputs. Our proposed module, SCON consists two normalization modules. Namely, they are **1.** a free-energy normalization module, and **2.** a energy-function output normalization module (*see* Fig. 1(b) for a schematic).

4.2.1 Free-energy Normalization for SCON

The objective of free-energy normalization is to encourage the model to produce outputs where the free-energy scores are normalized. We realize it as a training objective:

$$\mathcal{L}_{fen}^{(d)} = \frac{1}{N_d}\sum_{i=1}^{N_d} |F^{(d)}(x_i) - \tilde{F}^{(d)}(x_i)|. \tag{11}$$

Here, we compute the normalized free-energy, $\tilde{F}^{(d)}(x_i) = \big(F^{(d)}(\boldsymbol{x}_i) - \mu_{FE}^{(d)}\big)/\sigma_{FE}^{(d)}$ using global estimates for free-energy mean, $\mu_{FE}^{(d)}$ and variance, ${\sigma^2}_{FE}^{(d)}$. More formally, the normalization parameters are computed using the moving average of the mini-batch free energy scores as per,

$$\mu_{FE}^{(d)} = m \times \mu_{FE}^{(d)} + (1-m) \times \hat{\mu}_{FE}^{(d)}, \tag{12}$$

$${\sigma^2}_{FE}^{(d)} = m \times {\sigma^2}_{FE}^{(d)} + (1-m) \times \hat{\sigma^2}_{FE}^{(d)}. \tag{13}$$

Here, $\hat{\mu}_{FE}^{(d)}$ and $\hat{\sigma^2}_{FE}^{(d)}$ are the mean and variance of the free-energy scores computed on a mini-batch for a single domain, $d \in \{\text{“s”}, \text{“t”}\}$. For all our experiments we fix the momentum, m to 0.1. Note, the $\tilde{F}^{(d)}(x_i)$ computation here is similar to the normalization step of BatchNorm [13]. Hence, we realize it with the help of BatchNorm.

4.2.2 Energy Normalization for SCON

Similar to the free-energy normalization objective, $\mathcal{L}_{fen}$ we define our energy-score normalization objective, $\mathcal{L}_{en}$ as in,

$$\mathcal{L}_{en}^{(d)} = \frac{1}{N_d}\sum_{i=1}^{N_d} \|E(x_i^{(d)}) - \tilde{E}(x_i^{(d)})\|_2. \tag{14}$$

Note, that unlike the scalar free-energy scores, the energy-function output is a vector with a dimensionality equal to the number of classes.

The two normalizer modules are intended to work towards a similar objective *i.e.*, constrain the free-energy/energy scores. As such, we construct the overall normalization loss as a convex combination of two normalization losses (*see* Fig. 1(b) for a schematic). Thereby, the overall training loss for SCON can be written as in,

$$\mathcal{L}_{n}^{(d)} = \lambda \times \mathcal{L}_{fen}^{(d)} + (1-\lambda) \times \mathcal{L}_{en}^{(d)}. \tag{15}$$

Here, $\lambda < 1$ is a fixed positive scalar deciding the contribution of each normalization term. As shown in the schematic diagram (*see* Fig. 1(c)) we include two SCON modules for the source and the target, respectively. Thereby, the overall normalization loss for training is computed for both by summing the individual losses for both domains.

Overall training loss. Our overall training objective is the combination of the aforementioned training losses. As such, we write down our final training objective, $\mathcal{L}_T$ as in,

$$\mathcal{L}_T = \mathcal{L}_s + \alpha_u\mathcal{L}_u + \alpha_{ea}\mathcal{L}_{ea} + \alpha_n\left(\sum_{d\in\{\text{“s”},\text{“t”}\}} \mathcal{L}_n^{(d)}\right).$$

Here, the supervised loss, $\mathcal{L}_s = \frac{1}{N_s}\sum_{j=1}^{N_s} h_{ce}(\boldsymbol{x}_j^{(s)}, y_j^{(s)})$ is computed using the softmax cross-entropy, h_{ce} over labelled source instances. For all our experiments we keep, $\alpha_u = \alpha_{ea} = \alpha_n = 0.1$ constant.

5. Related Work

Energy-based learning. LeCun *et al.* [16], proposes energy-based learning as an alternative to probabilistic estimation for prediction, classification or decision-making tasks. The score based learning rules in energy-based learning are considered to be a more flexible way of implementing learning compared to probabilistic estimators.

The seminal work of Grathwohl *et al.*, [11], reinterpret classifiers as energy-based models. They propose a framework to learn the energy-based models simulating unlabelled data sampling from the marginal data distribution in the gradient space [35]. Note, this line of energy-based learning moves in the direction of generative modelling schemes. The energy-based learning we use in our work is more inline with the line of work discussed in [16] and is in a different scope to [11].

In [17], free-energy regularization is proposed as a stable form of regularization independent of pseudo labelling for UDA. Most related to our work is the energy-based active domain adaptation method proposed in [36]. They show

that domain invariant representations can be learnt by minimizing a disparity between source and target domain free-energy scores (*i.e.*, FEA). In this work, we show the importance of the proposed normalization, SCON to FEA. Furthermore, we show how free-energy can be used to do instance selection for self-training.

Distribution matching for UDA: Most UDA methods rely on domain distribution matching. Such algorithms can be separated as two schools of work. Namely, **1.** methods that explicitly align domain statistics [31], and **2.** domain-adversarial methods [4, 8, 19, 28]. Common to both these approaches is a minimization of a measure that estimates the domain disparity. The first school of methods relies on the alignment of explicit domain statistics such as covariance [31], MMD [3, 29]. Such methods often rely on prior assumptions to keep the statistical estimates tractable. For instance, [31] assumes the domain features are multivariate Gaussian distributions. In contrast, domain adversarial methods use a domain classifier to estimate the disparity between the domains. The domain-classifier is trained to discriminate instances from the domains. Thereafter, GAN [9] principles are used for learning domain invariant features along a two-player min-max learning framework. Such min-max learning methods require careful selection of hyper-parameters and training procedures (*e.g.*, label smoothing, transients for gradient-reversal) [8, 14, 26].

In comparison to above, we use the domain properties encoded by the free-energy scores of a model. In this way we avoid the need of assumptions on prior distributional properties as in domain alignment methods. Furthermore, our domain invariant learning method only requires a minimization over a free-energy scores alignment and a normalization. Thereby, we avoid training complications as in adversarial methods. Our work is mostly related to the active domain adaptation method of [36]. In fact, we make use of the FEA loss from [36]. However, we notice that the performance of FEA can be significantly improved when combined with the proposed normalization module, SCON.

Instance selection for self-training. Since the introduction of the "Π-model" [27], consistency regularization and pseudo-labelling solutions have seen improvements. However, the effectiveness of such methods is constrained by the quality of the pseudo-labelled predictions (*i.e.*, *confirmation-bias* [1]). To this end, FixMatch [30] shows that using a fixed threshold over the weakly augmented teacher model predictions is helpful to filter reliable data points for the self-training task.

To this end, recent UDA methods SENTRY [23] and PACMAC [24] has proposed instance selection criteria using prediction consistency and confidence. PACMAC [24] seeks meaningful attention-maps for their ViT model using this approach. However, these methods in general follow the assessment of the conditional distribution of predictions for their self-training task. In contrast to such methods, we propose an instance selection criterion to emulate the more informative joint distribution between data and predictions.

6. Experiments

We conduct experiments on DomainNet [21], OfficeHome [34] and VISDA2017 [22] datasets. For all experiments we follow the protocol explained in [24]. We provide more dataset details in the supplementary material.

6.1. Methods we compare with

We compare the performance of our method with state-of-the-art UDA methods. Here, we provide brief descriptions of them. **1. Source**: Supervised training on the labelled source domain instances. **2. CDAN** [19]: The conditional domain adversarial method using a discriminator model conditioned on classifier predictions. **3. SENTRY** [23]: A min-max self-training algorithm for improving the selection of confident pseudo-labelled target domain instances based on prediction consistency. **4. PACMAC** [24]: A consistency based self-training training method. PACMAC considers the prediction consistency with multiple versions of the ViT attention maps. It is a three staged UDA method including a unsupervised pretraining and a supervised finetuning stage. For our comparisons, we only considered the MAE [12] initialized versions. For fair comparisons, we report results for PACMAC by running their publicly available code in our servers[4].

We use the acronym **SEEBS** for our method to represent its **SE**lf-training method using **E**nergy-**B**ased instance **S**election strategy. As **SEEBS+** we report our overall methods performance including free-energy alignment and the proposed normalization modules, *i.e.*, **SCAL** + **SCON**.

6.2. Implementation details

For all our experiments we used VIT-Base [7] with 16×16 image patches. Starting from the official MAE [12] checkpoint pretrained using ImegNet1K [5], we follow the initialization procedure of PACMAC [24] (*i.e.*, in-domain pretraining using MAE followed by source domain finetuning) up until the adaptation phase[5]. We implement our method using the code-base of PACMAC. For all our experiments we use AdamW [20] optimizer with a learning rate of 2×10^{-4}. As per hyper-parameters, we find the normalizer module parameter $\lambda = 0.01$ (*see* equation (15)) works best for DomainNet and OfficeHome experiments. For all other experiments we use $\lambda = 0.1$. For all DomainNet and OfficeHome experiments we report results of our method after 300 epochs of training. In comparison, for the relatively large VISDA2017 we report results after 20 epochs.

[4] We indicate our evaluations by a $*$ mark in our results tables.

[5] Note, we only use the MAE pretrained models for our experiments.

6.3. Results

In Table 1, Table 2 and Table 3 we compare our method using top-1 accuracy. Here, we highlight the **best** and the second-best results. We notice that our proposed method, SEEBS+ outperforms all other methods in $18/24$ domain sets for both DomainNet [21] and OfficeHome [34]. It is interesting to notice that our proposed energy-based self-training method alone outperforms PACMAC in $16/24$ cases (We notice similar performance in 2 cases.). We attribute this to the proposed instance selection criteria.

We observe that our method works relatively better in difficult adaptation sets. For instance, for DomainNet, Cl2Pa appears to be the most challenging set reporting lowest Source only training performance (*i.e.*, 61.5%). For this set, we observe a 3% improvement in accuracy in SEEBS with PACMAC and a notable jump of 6.8% when SCON/SCAL modules are included. We relate the improvement obtained by SEEBS in this case to the enhanced capacity of our instance selection criterion for self-training. On top of prediction consistency based instance selection, SEEBS probes into the joint distribution of the data for cleaner self-training signal. For instance, for domain sets where there is a significant shift in the covariates, our energy-based selection criterion appears to be most useful. We notice similar jumps in performance for other difficult domain sets (*e.g.*, Ar2Cl, Pr2Cl in OfficeHome dataset).

To further establish our methods potential and its scalability, in Table 3 we provide our comparisons on the challenging VISDA2017 [22] dataset. Here, we observe that both our SEEBS and SEEBS+ methods outperform PACMAC by a considerable margin (*i.e.*, $+2\%$ in accuracy).

6.4. Further analysis

6.4.1 Impact of self-training instance selection based on the joint distribution.

We propose a self-training method emulating the instance selection by considering the joint distribution of predictions and unlabelled target domain data-points. Here, in Table 4 we show its importance by comparing the performance when instance selection is done only considering the conditional distribution $P(y|x)$, given as **Cond.**. We realize this case by only using the self-training method proposed in PACMAC [24][6]. PACMAC considers consistency in predictions to identify reliable instance for self-training. As per the marginal distribution scenario, **Marg.**, it is equivalent to the case where we only consider the free-energy based selection criteria given in equation 7. To this end, we realize the joint distribution based instance selection **Joint.** by fusing the instance selections from **Cond.** and **Marg.**.As could be seen from Table 4 the proposed **Joint.** distribution based instance selection outperforms the other two methods in both the ablation cases. In Fig. 3 we compare the propotion of instances selected for PACMAC and SEEBS+ (for Rw2Ar). We provide a detailed qualitative analysis in the supplementary.

[6]For fair comparison, we run our ablation experiments for 300 epochs for all methods. For this reason, the reported results, **Marg.** may not match with PACMAC results reported in Table 1 and Table 2.

6.4.2 Impact of the low free-energy based instance selection condition.

Here we conducted additional evaluations to examine the impact of using instances that meet the free-energy based instance selection condition explained in equation (7). The method reported as **Sel-low** is the proposed free-energy based selection *see* Table 5 **(Left)**. As per, equation (4), this selection criterion evaluates good attention parameters by considering a lower free-energy than the corresponding non-attention masked input. To this end, we compare the choice of **Sel-high** where the instance selection rule is inverted. In the bottom two rows of the table we show how the instance selections for these two conditions change when used to form the joint distribution based selections (*i.e.*, by fusing with PACMAC). It is interesting to see that the inverted criterion **Sel-high** performs better in this marginal distribution case. However, it can be seen that when used to form the Joint-sel case, our proposed criterion gets better performance. We provide the expanded domain set results in the supplementary. We use this observation to establish our choice for the energy based instance selection.

6.4.3 Impact of SCON on free-energy alignment.

In Table 5 **(Right)**, we present the impact of our SCON module. SCAL is responsible for domain alignment, while SCON is a normalization process aimed at improving the stability of domain alignment. Our results show that SCAL positively impacts performance (*see* supplementary for the expanded table). Note, the reported method as OnlySCAL is equivalent to the FEA of [36]. However, when we combine SCAL with the proposed normalization process, we observe a significant improvement in performance.

6.4.4 Comparisons with naive-normalization

An alternative to our proposed normalization, **SCON** is the case where it is naively replaced by a BatchNorm [13]. We call this baseline as **naiveSEEBS+**. Note, in distinctive to **naiveSEEBS+**, the proposed **SCON** provides a training objective for the model to learn to produce normalized energy-scores. In Table 6 we compare **SEEBS+** with **naiveSEEBS+** (*see* supplementary for the full table). We observe that in **11/12** sets of DomainNet and in **8/12** sets of OfficeHome, **naiveSEEBS+** is outperformed by **SEEBS+**.

Method	Re2Cl	Re2Pa	Re2Sk	Cl2Re	Cl2Pa	Cl2Sk	Pa2Re	Pa2Cl	Pa2Sk	Sk2Re	Sk2Cl	Sk2Pa	AVG
Source	71.0	77.6	62.9	73.7	61.5	63.3	82.4	63.1	66.1	76.6	71.9	69.6	**70.1**
CDAN [19]	72.2	74.5	59.3	80.6	57.3	59.2	78.5	57.4	61.2	81.4	73.2	69.4	**67.7**
SENTRY [23]	84.2	82.8	76.4	86.9	77.1	74.1	86.9	76.2	73.3	**88.8**	81.6	77.6	**80.5**
PACMAC* [24]	86.5	82.2	78.3	84.9	72.5	75.8	**88.6**	**84.1**	79.2	84.6	83.0	78.8	**81.5**
SEEBS	87.1	82.9	78.3	85.5	75.5	76.6	87.8	82.7	78.1	82.5	82.7	78.0	**81.5**
SEEBS+	**90.0**	**83.8**	**80.2**	**87.2**	**79.3**	**78.3**	88.1	83.9	**79.8**	84.6	**84.5**	**80.6**	**83.3**

Table 1. **Comparisons on DomainNet [21].** Here we compare the results of our proposed method **SEEBS+** with SOTA UDA methods. **SEEBS** is the version of our method where we exclude the SCAL and SCON modules. Our method gives top performance in $9/12$ sets.

Method	Ar2Cl	Ar2Pr	Ar2Rw	Cl2Ar	Cl2Pr	Cl2Rw	Pr2Ar	Pr2Cl	Pr2Rw	Rw2Ar	Rw2Cl	Rw2Pr	AVG
Source	46.7	57.6	71.0	51.1	60.0	62.6	51.4	46.9	70.5	66.3	52.2	77.2	**59.4**
CDAN [19]	45.3	58.8	69.1	51.6	60.7	61.5	53.4	45.5	72.4	67.7	49.9	78.0	**59.5**
SENTRY [23]	54.8	65.6	74.4	56.5	65.8	**69.8**	57.6	54.9	75.5	68.9	60.0	81.6	**65.5**
PACMAC* [24]	58.2	64.3	73.5	60.0	68.7	67.0	57.0	57.9	74.0	69.8	63.6	81.8	**66.3**
SEEBS	**61.3**	65.2	74.0	60.4	**69.6**	67.3	57.2	59.1	73.9	70.9	64.7	**82.0**	**67.1**
SEEBS+	60.9	**68.1**	**75.3**	**62.1**	68.7	68.1	**60.6**	**60.0**	**75.7**	**72.2**	**68.2**	**82.0**	**68.5**

Table 2. **Comparisons on OfficeHome [34].** Here we compare the results of our proposed method **SEEBS+** with SOTA UDA methods. **SEEBS** is the version of our method where we exclude the SCAL and SCON modules. Our method gives top performance in $11/12$ sets.

Method	Acc%
Source	78.6
PACMAC* [24]	77.1
SEEBS	79.1
SEEBS+	**79.5**

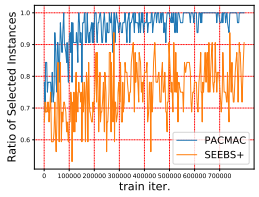

Table 3. **(Left) VISDA2017.** Here we compare the results of our proposed method **SEEBS+** with SOTA UDA methods. **SEEBS** is the version of our method without SCAL and SCON. **(Right)** The proportion of instance selections for PACMAC and SEEBS+.

Method	DomainNet	OfficeHome
Cond.	80.9	66.8
Marg.	80.7	66.1
Joint.	**81.5**	**67.1**

Table 4. Comparison of self-training instance selection based on the conditional, marginal, and the joint distribution. Here we report the average performance for each case across all domain sets. An expanded table is provided in the supplementary.

7. Conclusions

We proposed a UDA method by making use of EBL. Namely, we provided details of two directions on how UDA can benefit from EBL, **1.** EBL can be used to improve the instance selection for a self-training task on the unlabelled target domain, and **2.** alignment and normalizing energy scores can learn domain-invariant representations.

Method	DomainNet	OfficeHome
Sel-high	**81.1**	**66.4**
Sel-low	80.7	66.1
Joint-sel-high	80.1	66.6
Joint-sel-low	**81.5**	**67.1**

Method	DomainNet	OfficeHome
Source	70.1	59.4
OnlySCAL	77.2	63.2
OnlySCON	46.4	50.9
SCAL+SCON	**82.1**	**68.2**

Table 5. **(Left)** The impact of using instances that meet the proposed free-energy based instance selection condition in (7). **(Right)** The effectiveness of the combined SCAL and SCON modules. Here we report the average performance for each case across all domain sets. We provide expanded tables in the supplementary.

	DomainNet	OfficeHome
naiveSEEBS+	81.4	68.0
SEEBS+	**83.3**	**68.5**

Table 6. A comparison of SEEBS+ to the case where the proposed SCON module is replaced with a BatchNorm [13], naiveSEEBS+. Here we report the average performance for each case across all domain sets. We provide the expanded table in the supplementary.

We realize our proposed method benefiting from the attention mechanism in ViTs. To establish our claims we show that our method outperforms state-of-the-art UDA methods in $19/25$ domain transfers. Lastly, we provide ablations to justify the significance of the proposed method.

Acknowledgement. This work was supported by an eBay Research Award, and DARPA's Learning with Less Labeling (LwLL) under agrement FA8750-19-2-0501. We acknowledge the support from the Australian Research Council (ARC) for M. Harandi's project DP230101176, and B. Fernando's NRF Fellowship (NRF-NRFF14-2022-0001) from the National Research Foundation, Singapore.

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2020. 6

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

[3] Mahsa Baktashmotlagh, Mehrtash T Harandi, Brian C Lovell, and Mathieu Salzmann. Unsupervised domain adaptation by domain invariant projection. In *Proc. Int. Conference on Computer Vision (ICCV)*, pages 769–776, 2013. 6

[4] Zhangjie Cao, Lijia Ma, Mingsheng Long, and Jianmin Wang. Partial adversarial domain adaptation. In *Proc. European Conference on Computer Vision (ECCV)*, pages 135–150, 2018. 6

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

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, 2019. 1

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. Int. Conference on Learning Representations (ICLR)*, 2021. 1, 6

[8] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proc. Int. Conference on Machine Learning (ICML)*, pages 1180–1189, 2015. 6

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 6

[10] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Proc. Advances in Neural Information Processing Systems*, 2004. 1

[11] Will Grathwohl, Kuan-Chieh Wang, Joern-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 *Proc. Int. Conference on Learning Representations (ICLR)*, 2020. 1, 4, 5

[12] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 6

[13] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. Int. Conference on Machine Learning (ICML)*, pages 448–456, 2015. 5, 7, 8

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134, 2017. 6

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. Advances in Neural Information Processing Systems*, 2012. 5

[16] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006. 1, 5

[17] Xiaofeng Liu, Bo Hu, Xiongchang Liu, Jun Lu, Jane You, and Lingsheng Kong. Energy-constrained self-training for unsupervised domain adaptation. In *Proc. Int. Conference on Pattern Recognition (ICPR)*, pages 7515–7520, 2021. 5

[18] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *Proc. Int. Conference on Learning Representations (ICLR)*, 2021. 3

[19] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2018. 6, 8

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[21] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proc. Int. Conference on Computer Vision (ICCV)*, pages 1406–1415, 2019. 2, 6, 7, 8

[22] 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. 2, 6, 7

[23] Viraj Prabhu, Shivam Khare, Deeksha Kartik, and Judy Hoffman. Sentry: Selective entropy optimization via committee consistency for unsupervised domain adaptation. In *Proc. Int. Conference on Computer Vision (ICCV)*, pages 8558–8567, 2021. 1, 3, 6, 8

[24] Viraj Prabhu, Sriram Yenamandra, Aaditya Singh, and Judy Hoffman. Adapting self-supervised vision transformers by probing attention-conditioned masking consistency. In *Proc. Neural Information Processing Systems (NeurIPS)*, pages 23271–23283, 2022. 1, 3, 4, 6, 7, 8

[25] Yunchen Pu, Shuyang Dai, Zhe Gan, Weiyao Wang, Guoyin Wang, Yizhe Zhang, Ricardo Henao, and Lawrence Carin Duke. Jointgan: Multi-domain joint distribution learning with generative adversarial nets. In *Proc. Int. Conference on Machine Learning (ICML)*, pages 4151–4160, 2018. 1

[26] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 6

[27] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In *Proc. Advances in Neural Information Processing Systems*, pages 3546–3554, 2015. 6

[28] Rui Shu, Hung Bui, Hirokazu Narui, and Stefano Ermon. A DIRT-t approach to unsupervised domain adaptation. In

Proc. Int. Conference on Learning Representations (ICLR), 2018. 1, 6

[29] Alexander J Smola, A Gretton, and K Borgwardt. Maximum mean discrepancy. In *13th international conference, ICONIP*, 2006. 6

[30] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Proc. Neural Information Processing Systems (NeurIPS)*, pages 596–608, 2020. 3, 6

[31] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Proc. European Conference on Computer Vision (ECCV)*, pages 443–450, 2016. 6

[32] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 5

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 1

[34] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5018–5027, 2017. 2, 6, 7, 8

[35] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011. 5

[36] Binhui Xie, Longhui Yuan, Shuang Li, Chi Harold Liu, Xinjing Cheng, and Guoren Wang. Active learning for domain adaptation: An energy-based approach. In *Proc. of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 8708–8716, 2022. 1, 2, 4, 5, 6, 7

[37] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proc. Int. Conference on Computer Vision (ICCV)*, pages 2849–2857, 2017. 1

[38] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. Int. Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017. 1