SCALE: Semantic Contrast-Aware Learning for Domain-Adaptive Image Classification

Zhuoran Li **Yiming Ma** **Qi Zhang**

Jianzhi Teng **Yilin Wu**

Hong Kong Polytechnic University
{23046058g, 23050087g, 23065105g, 23067237g, 23059361g, 23060452g}@connect.polyu.hk

Abstract

Domain adaptive image classification aims to generalize the knowledge learned from labeled known domains to label-free, unforeseen domains through a supervised learning algorithm. The core of such a transfer learning method is to learn the domain shift between the source and target domain, which might be challenging due to the extreme mismatch of two distributions. This domain shift can significantly degrade the performance of a model trained thoroughly on the source domain data when applied to the target domain. Therefore, domain adaptive learning algorithms for image classification are expected to be proposed to extensively generalize the knowledge distilled from the given image-level distribution without requiring prior labeling or retraining on the target domain. Inspired by semantic segmentation-based domain transfer learning, we propose SCALE, a novel image classification adaptation framework that learns semantic-instructed contrast via a teacher-student architecture for image representation across underlying domains. This approach is capable of learning a cross-domain discriminative pixel feature representation by utilizing different semantic concepts to guide pixel-level comparison learning, thus improving the performance of self-training. We perform extensive experiments on downstream classification tasks to validate the effectiveness of our presented method and further demonstrate the efficiency of such a self-training process. The proposed approach is theoretically and empirically verified to be feasible in discriminative representation learning as well as showcasing excellent scalability on flexibly using disparate forms of semantic concept of the image.The video presentation is released here

1 Introduction

Real-world data exhibiting mismatch or domain shift [1] poses an indispensable challenge to the generalization capability of machine learning model construction. Different from the ideal training data, the real-scenario data has higher complexity and randomness across the distribution. Applying a model's learned data distribution to a wider range of test data often results in drastic degradation of the model performance[2]. Due to various factors such as changes in lighting conditions, viewpoints, backgrounds, or data acquisition processes, the distribution among images from the distribution displays a significant difference in general. This poses several challenges for the usage of pretrained models in downstream image classification tasks[10]. Besides this, motivations such as the lack of labeled data, model deployment in the diverse production environment, fairness and bias mitigation necessity, as well as computational efficiency requirements have put forward high standards for model generalization capability, and it is expected that a unified framework of modeling can be demonstrated to achieve effective knowledge transfer between domains through small-sample, good-performance learning[4, 9, 19, 20]. As a consequence, the domain adaptive techniques were developed to bridge the gap between the source and unforeseen domain, i.e. the gap across image-level distribution in the proposed context. By handling the difficulty caused by domain shifts and leveraging knowledge from source domains, domain adaptive methods for image classification can improve model performance and robustness in practical deployment scenarios[6]. They enable more efficient use of limited labeled data and facilitate the adaptation of models to diverse real-world environments.

The proposed approach is a broad extension of semantic contrast-aware learning, an advanced technique used in transfer image classification learning tasks that focuses on capturing and leveraging semantic contrast information to improve the accuracy and discriminative power of the classification model[16, 25]. In traditional image classification, models typically rely on low-level features such as color[22], texture, and shape to distinguish between different classes. However, these features may not always capture the underlying semantic differences between objects or regions in an image. SCALE addresses this limitation by incorporating semantic contrast information into the learning process[24, 30]. It aims to learn the subtle differences and relationships between different semantic regions in an image, enabling the

All authors contributed equally.
The code is released at https://github.com/wellssssss/SCALE.

model to achieve more informed and accurate classification performance.

SCALE involves several key procedures: semantic segmentation which includes semantical division, understanding, and labeling; contrastive learning which learns representations that bring similar instances closer together while pushing dissimilar instances apart[10]. It leverages pairs of augmented versions of the same image and encourages the model to capture the underlying semantic contrasts; feature extraction that includes meaningful feature selection and semantics encoding[12, 29, 8]; and image classification via deep neural network to map the extracted features to specific class labels. By incorporating semantic contrast information into the learning process, SCALE can effectively capture the subtle differences between different semantic regions within an image. This improves accuracy and robustness in image classification tasks, especially when dealing with complex scenes or fine-grained classification problems[21].

We use a teacher-student architecture to further improve the model performance in the scenario of the self-training[23] process. This architecture plays a crucial role in the self-training process and the overall learning of cross-domain discriminative pixel feature representations. Our teacher-student architecture consists of two models: the teacher model and the student model. Both models share a similar structure, comprising an encoder a classification module, and an auxiliary projection head[5]. The teacher model is responsible for generating pseudo-labels for the target domain data. During the training process, the input images from both the source and target domains are processed by the teacher model[14, 19]. The student model is the model being trained and optimized during the domain adaptation process. Like the teacher model, the student model processes the input images from both the source and target domains. In a nutshell, by utilizing the teacher-student architecture, the SCALE framework can effectively leverage the available labeled source domain data and unlabeled target domain data to learn cross-domain discriminative pixel feature representations, ultimately improving the performance of domain adaptive image classification[28].

Consequently, such a proposed transfer learning framework with a scalable domain adaptation capability to overcome the distribution gap across images. We demonstrate an end-to-end form domain adaptive image classifier, SCALE, theoretically, whose core mechanism is to learn a cross-domain discriminative pixel feature representation by utilizing different semantic concepts to guide the pixel-level contrastive learning process, thereby improving the performance of self-training-based domain adaptation[23, 30]. We adopt a teacher-student architecture, where both the teacher model and the student model comprise an encoder, a classification module, and an auxiliary projection head. The input images are processed by the teacher and student models, with the student model generating source and target domain features,

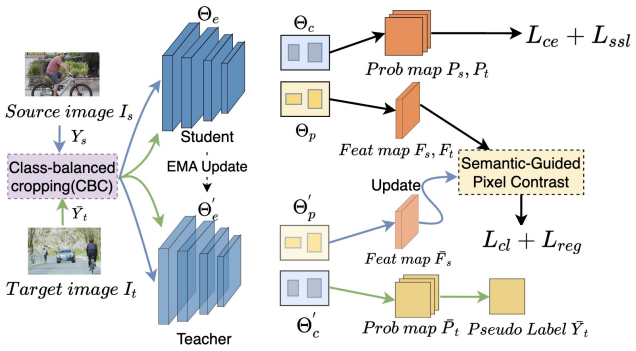

Figure 1: Framework overview. First, the method is based on a teacher-student architecture, where the teacher model provides source feature maps and target pseudo-labels. Secondly, class-balanced cropping is proposed to frequently crop image patches containing rare class samples to balance the performance of different classes. Furthermore, in addition to the self-training loss, a contrastive loss is also used to align the image feature representations towards the cluster centers. Finally, after training is completed, the projection head is discarded and only the encoder and classification head are used for image classification tasks.

i.e. source feature and target features as well as corresponding classification results, i.e. source logits and target logits, while the teacher model produces pseudo-labels for the target domain data[30].

Substantial experiments are conducted to verify the feasibility of the proposed method. We use Office-Home as the benchmark dataset to examine the transfer learning capability of the proposed method. We train and finetune a ResNet-based image classifier to fit into the SCALE framework. The model achieves an average classification accuracy of 75.09% which is only approximately 3% lower than that of the SOTA approach, indicating the superiority and robustness of our method. Meanwhile, SCALE has its model performance far exceeding the baseline method, empirically expressing the success of the domain transfer learning capability of SCALE. In short, our contribution can be summarized as

Our Contribution (1) We propose a novel transfer learning framework to solve domain mismatch in image classification by utilizing different semantic concepts to guide pixel-level comparison learning, a cross-domain discriminative pixel feature representation method, which improves the performance of self-training. (2) We present the teacher-student architecture that successively performs aggregation and representational semantics learning in unsupervised learning. This aggregated architecture significantly improves computational efficiency compared to traditional ensemble transfer learning algorithms. (3) By applying self-training, supervised source domain, and pseudo-label self-training target domain loss, we learn a more comprehensive semantic guidance for contrast learning. We further prove that this distribution-aware optimization algorithm is theoretically effective in enhancing the model performance. (4) Extensive experiments are performed

to demonstrate the empirical effectiveness of the proposed approach. SCALE achieves an average classification accuracy of 75.09%, which is only lower than SOTA by 3.11%.

2 Related Work

Domain adaptation is a crucial research direction in computer vision that focuses on addressing the distribution discrepancy between the source domain and the target domain. Over the years, various approaches have been proposed to tackle this problem. [1] analyzes the trade-offs involved in designing a representation for domain adaptation and provide a new justification for a recently proposed model. [2] introduces Margin Disparity Discrepancy and develop an adversarial learning algorithm for domain adaptation based on their theoretical framework. [3] introduces a new representation learning approach for domain adaptation, in which data at training and test time come from similar but different distributions.[4] proposes a new approach to domain adaptation in deep architectures that can be trained on large amount of labeled data from the source domain and large amount of unlabeled data from the target domain (no labeled target-domain data is necessary).

In recent years, there have been further advancements in domain adaptation research. [5] proposes Learning to Match (L2M), an approach that automatically learns cross-domain distribution matching without relying on hand-crafted priors for the matching loss. [6] takes a different approach to the problem of generalizing across data sources. [7] proposes a simple yet effective enhancement for Mixup-based domain generalization (DG) and introduce the domain-invariant Feature MIXup with Enhanced Discrimination (FIXED) approach, which utilizes existing techniques to enlarge margins between classes.

These works highlight the ongoing efforts in developing novel methods and approaches to improve domain adaptation in computer vision, enabling models to generalize better across different domains and enhance the overall performance of real-world applications.

Knowledge Distillation has gained significant attention in the field of machine learning and has been extensively studied by researchers. One of the pioneering works in knowledge distillation was introduced by [8] in 2015. It significantly improve the acoustic model of a heavily used commercial system by distilling the knowledge in an ensemble of models into a single model. [9] in 2014 directly extracted semantic information from the intermediate features of the teacher model and compressing a wide and deep teacher model into a thin and deep student model can lead to better generalization and faster execution. divided the distillation into two parts: Soft distillation and Hard-label distillation. Moreover they also introduce a distillation token to reproduce the teacher model's predicted labels during knowledge distillation.

In recent years, several advancements and novel approaches have been proposed in the field of knowledge distillation, pushing the boundaries of its applications and effectiveness. These works have focused on various aspects, including: [10] divides distillation into two kind of approaches: Soft distillation and Hard-label distillation and also introduce a distillation token is a technique used to reproduce the teacher model's predicted labels during knowledge distillation. [11] decouples representation learning and classification, the LFM loss considers feature representation, while the LSR loss ensures that the teacher and student features produce the same output when passed through the teacher's frozen classifier for the same input image. [12] decouples classical distillation into two components: Target Class Knowledge Distillation (TCKD) and Non-Target Class Knowledge Distillation (NCKD). This decoupling enhances the effectiveness and flexibility of knowledge transfer. [13] introduces of a self-supervised teaching assistant (SSTA) is incorporated to guide the learning of the student model along with a supervised teacher (SLT). [14] investigates the potential of knowledge distillation from a pretrained Masked Autoencoders (MAE) model by aligning the intermediate features between a larger MAE teacher model and a smaller MAE student model.

3 Background

We demonstrate the domain adaptive image classification by applying the model of encoders, self-training domain adaptation and pixel contrast methods by comparing the theoretical basis. We formulate the background of the problem in this section[4, 21, 25].

3.1 Encoder

An encoder[7, 9] is a neural network model designed to learn a compressed representation of the input. They are an unsupervised learning method, although technically they are trained using a supervised learning method called self-supervision. Encoders are typically trained as part of a broader model that attempts to recreate the input.

For example:

$$X = \text{model.predict}(X)$$

This challenge is made challenging by the design of the encoder model intentionally constraining the architecture to a bottleneck at the midpoint of the model from which reconstruction of the input data is performed[11]. In this paper, we use it as a learning or automatic feature extraction model to embedding images, and ultimately achieve image classification.

The encoder encodes an input by mapping it to a hidden representation through a deterministic nonlinear function, Specifically, for a given data set $x^{(i)} \in R^d, i = 1 \cdots N$,the encoder encodes x^i through $z^{(i)} = f(W_e x^i + b)$. Here, $f(\cdot)$is a component wise nonlinear activation function. $W_e \in R^{s \times d}$ and $b \in R^s$ are the encoder weight matrix and bias, respectively. The encoder encourages $\tilde{x}^{(i)}$ to be as close as possible to $x^{(i)}$ under a given distance metric.

3.2 Domain Adaptation

Domain Adaptation is an important research area[13, 14, 16] that aims to address the problem of different distributions between training data (source domain) and test data (target domain).

Feature Representation-based Methods: The goal of these methods is to learn a shared feature representation space where the distributions of the source and target domains are as close as possible. Representative methods include:

Maximum Mean Discrepancy (MMD): MMD measures the mean difference between the source and target domains in a Reproducing Kernel Hilbert Space (RKHS)[15]. By minimizing MMD, domain-invariant features can be learned.Among them, $\phi(\cdot)$ represents the mapping function, and $\mathcal{H}$ represents the Hilbert space.

$$\text{MMD}(P, Q) = \left\| \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) - \frac{1}{m} \sum_{j=1}^{m} \phi(y_j) \right\|_{\mathcal{H}}^2$$

Deep Adaptation Networks (DAN): DAN utilizes the MMD [20] measure at multiple layers of a deep neural network to align the feature distributions of the source and target domains.Where $\mathcal{F}^{(l)}S$ and $\mathcal{F}_T^{(l)}$ represent the feature representation of the source domain and the target domain in the lth layer respectively.

$$\mathcal{L}_{\text{DAN}} = \sum_{l=1}^{L} \text{MMD}(\mathcal{F}_S^{(l)}, \mathcal{F}_T^{(l)})$$

Adversarial Learning-based Methods: Inspired by Generative Adversarial Networks (GANs), these methods introduce a domain discriminator[20, 21] to distinguish between source and target domain data, while training a feature extractor to fool the domain discriminator[16]. Through adversarial training, the feature extractor can learn domain-invariant feature representations. Representative methods include: Domain-Adversarial Neural Networks (DANN): DANN introduces a domain discriminator and optimizes the feature extractor by backpropagating gradients to learn domain-invariant features. Among which, $\mathcal{L}$cls represents the classification loss, $\mathcal{L}$adv represents the adversarial loss, and λ is the loss weight.

$$\mathcal{L}_{\text{DANN}} = \frac{1}{n} \sum_{i=1}^{n} \left(\mathcal{L}_{\text{cls}}(x_i, y_i) - \lambda \mathcal{L}_{\text{adv}}(x_i) \right)$$

Conditional Domain Adversarial Networks (CDAN): CDAN extends DANN by considering class information and employs a conditional domain discriminator to improve the effectiveness of adversarial learning.

Self-training-based Methods: These methods utilize the label information from the source domain to predict pseudo-labels for the target domain data, and then use these pseudo-labels to train the model[17]. Through multiple iterations, the model can gradually adapt to the target domain data distribution. Like: Mean Teacher: The Mean Teacher method maintains a teacher model and a student model, where the teacher model's parameters are an exponential moving average of the student model's parameters. By minimizing the prediction difference between the teacher and student models on the target domain, more stable pseudo-labels can be obtained.In which, Among them, $\mathcal{F}$teacher(x_i) and $\mathcal{F}$student(x_i) represent the feature of sample x_i in the teacher model and student model respectively.

$$\mathcal{L}_{\text{MT}} = \frac{1}{n} \sum_{i=1}^{n} \text{MSE}(\mathcal{F}_{\text{teacher}}(x_i), \mathcal{F}_{\text{student}}(x_i))$$

Self-ensemble Learning: Self-ensemble learning methods generate multiple views by applying perturbations to the model's outputs and inputs. The predictions from these views are then ensembled to obtain more robust pseudo-labels.

3.3 Image Contrast Learning

Image contrast methods [20] are a class of techniques used for unsupervised visual representation learning. These methods aim to learn image-level representations by maximizing the similarity between different views of the same image while minimizing the similarity between different images. Here are several main image contrast methods:

Momentum Contrast for Unsupervised Visual Representation Learning: MoCo[20, 21, 22, 23] is a powerful image contrast method. It maintains a dynamic dictionary of negative samples, which are encoded by a momentum-updated encoder. The momentum encoder is the slowly moving average of the base encoder, which provides consistent representations for the negative samples. For each input image, MoCo generates a query encoding using the base encoder and a key encoding using the momentum encoder: $\theta_k \leftarrow m\theta_k + (1-m)\theta_q$, where θ_k denotes the parameters of the momentum encoder (key encoder), θ_q represents the parameters of the query encoder, and $m \in [0, 1)$ is the momentum coefficient. The query is then compared against the keys in the dictionary using a contrastive loss. By treating the query's matching key as a positive sample and all other keys as negative samples, MoCo learns image-level representations that are discriminative and robust.

Semantic Pixel Contrast: SePiCo is an approach [27] that aims to learn discriminative pixel-level feature representations by leveraging contrastive learning. The core idea is to encourage pixels belonging to the same semantic class to have similar feature representations while pushing apart the representations of pixels from different classes. The method introduces a pixel-level contrastive loss $\mathcal{L}cl$ that maximizes the similarity between each pixel's feature vector z_i and its corresponding class prototype vector cy_i, while minimizing the similarity with prototype vectors of other classes. The prototype vector c_k for class k is computed as the average of feature

vectors of all pixels belonging to that class:

$$c_k \leftarrow \frac{\sum_{i=1}^{N} \mathbb{1}[y_i = k] z_i}{\sum_{i=1}^{N} \mathbb{1}[y_i = k]}$$

where $\mathbb{1}[y_i = k]$ is an indicator function that equals 1 when the class label y_i of the i-th pixel is k, and 0 otherwise. N is the total number of pixels. By updating the prototype vectors using the average of pixel feature vectors within each class, SePiCo effectively captures the semantic information and learns a more compact and discriminative feature space. The contrastive loss encourages pixels of the same class to cluster together around their corresponding prototype vector, while pushing apart clusters of different classes.

4 Proposed Method

We first briefly introduce the architecture of our image classification model. Then the demonstration of encoders and self-training domain adaptation are in Section 4.2 and Section 4.3. Pixel contrast method is elaborated in Section 4.4. We finally demonstrate the training processing.

4.1 Teacher Student Architecture

The teacher network takes a source image $I_s \in \mathbb{R}^{H \times W \times 3}$ as input and generates a source feature map $\bar{F}_s \in \mathbb{R}^{H' \times W' \times A}$ and target pseudo labels $\bar{Y}_t \in \mathbb{R}^{H \times W \times K}$. To address class imbalance, we introduce a class-balanced cropping (CBC) strategy that frequently crops image patches containing under-represented objects, aiming to balance performance across different classes.

During the training process, images from both the source domain I_s and target domain I_t, with dimensions $H \times W \times 3$, are randomly sampled and fed into the student and teacher networks simultaneously. The student network processes these images and generates hidden-layer features $F_s, F_t \in \mathbb{R}^{H' \times W' \times A}$ and final pixel-level predictions $P_s, P_t \in \mathbb{R}^{H \times W \times K}$. Here, A represents the number of channels in the intermediate features, while H' and W' denote the spatial dimensions of the features, which are typically much smaller than the input image dimensions H and W. K stands for the number of classes in the segmentation task.

Correspondingly, we obtain the source features $\bar{F}_s \in \mathbb{R}^{H' \times W' \times A}$ and target pixel-level predictions $\bar{P}_t \in \mathbb{R}^{H \times W \times K}$ from the teacher network, which is updated using a momentum-based strategy. The teacher network's parameters $\Theta' t$ are updated as a weighted average of the student network's parameters Θ_t and the previous teacher network's parameters $\Theta' t - 1$, as shown in the equation $\Theta' t = \alpha \Theta' t - 1 + (1 - \alpha)\Theta_t$, where α is the momentum coefficient.

4.2 Encoder Architecture

We have opted to employ ResNet as the encoder architecture in our proposed framework. The loss function used in ResNet is the same as that in common convolutional neural networks, typically employing the Cross-Entropy Loss for classification tasks. For a batch of N samples, the Cross-Entropy Loss is defined as:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log(\hat{y}_{i,c})$$

where $y_{i,c}$ is the true label (0 or 1) indicating whether the i-th sample belongs to class c, $\hat{y}_{i,c}$ is the predicted probability of the i-th sample belonging to class c, and C is the total number of classes.

In ResNet, assuming $\mathbf{x}$ is the input feature and $\mathcal{F}(\mathbf{x})$ represents the residual function, the output of a residual unit is:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}) + \mathbf{x}$$

where $\mathcal{F}(\mathbf{x})$ is typically composed of two or three convolutional layers, with a ReLU activation function following the last convolutional layer.

During back-propagation, let $\frac{\partial \mathcal{L}}{\partial \mathbf{y}}$ be the gradient of the loss function with respect to the output of the residual unit. Then, according to the chain rule, the gradient of the loss function with respect to the input feature $\mathbf{x}$ is:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{x}} = \frac{\partial \mathcal{L}}{\partial \mathbf{y}} \cdot \frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \frac{\partial \mathcal{L}}{\partial \mathbf{y}} \cdot (1 + \frac{\partial \mathcal{F}(\mathbf{x})}{\partial \mathbf{x}})$$

Due to the presence of the identity mapping, even if $\frac{\partial \mathcal{F}(\mathbf{x})}{\partial \mathbf{x}}$ is very small or zero, the gradient $\frac{\partial \mathcal{L}}{\partial \mathbf{x}}$ can still be propagated to earlier layers through the identity mapping, thus alleviating the vanishing gradient problem.

4.3 Self-training Domain Adaptation

The main idea is to use the model's high-confidence predictions on the target domain to generate pseudo-labels, which are then used to retrain the model. In the self-training process, the loss function plays a crucial role. The goal of self-training is to improve the model's performance on the target domain by leveraging unlabeled data from the target domain. In the self-training process for domain adaptation, the loss function guides the model to learn from both the labeled source domain data $\mathcal{D}_S = (\mathbf{x}i^S, y_i^S)i = 1^{n_S}$ and the unlabeled target domain data $\mathcal{D}_T = (\mathbf{x}i^T)i = 1^{n_T}$[26]. The model is initially trained on the source domain using cross-entropy loss:

$$\mathcal{L}ce = -\frac{1}{n_S} \sum_{i=1}^{n_S} \sum_{j=1}^{C} y_{ij}^S \log f_\theta(\mathbf{x}_i^S)_j$$

where y_{ij}^S is the true label of the source domain sample $\mathbf{x}i^S$, and $f\theta(\mathbf{x}_i^S)_j$ is the predicted probability for the j-th class by

the model. The trained model is then used to make predictions on the target domain data, and high-confidence predictions are selected as pseudo-labels $\hat{y}i^T i = 1^{n_T}$. These pseudo-labeled target domain samples are added to the training set, and the model is retrained using a combination of the supervised loss and a consistency loss:

$$\mathcal{L} = \mathcal{L}ce + \lambda \mathcal{L}cons$$

where $\mathcal{L}_{cons}$ is the consistency loss, which encourages the model to make consistent predictions on perturbed versions of the target domain samples:

$$\mathcal{L}cons = \frac{1}{n_T} \sum {}^{--} i = 1^{n_T} |f_\theta(\mathbf{x}i^T) - f\theta(\mathbf{x}_i^{T,perturb})|^2$$

Here, $\mathbf{x}_i^{T,perturb}$ is the result of applying random perturbations to the target domain sample $\mathbf{x}_i^T$, and λ is a weight to balance the supervised loss and consistency loss. By iteratively refining the model using the pseudo-labeled target domain samples and the consistency loss, the model adapts to the target domain data distribution and improves its performance on the target domain[22]. This process continues until a stopping criterion is met.

4.4 Pixel Contrast Learning

Inspired by the work of [30], we consider adapting their pixel contrast methods to our image domain adaptation task. In the literature, a unified framework was designed to integrate three distinct contrastive losses that target learning similar/dissimilar pairs at the pixel level to mitigate the domain gap via either centroid-aware pixel contrast or distribution-aware pixel contrast[30]. Moreover, to encourage globally diverse and smooth feature representations of the input images, [30] introduce a regularization term defined as:

$$L_{reg} = -\frac{1}{K} \sum_{k=1}^{K} \log \frac{\exp(\bar{Q}^\top \mu_k/\tau)}{\sum_{l=1}^{K} \exp(\bar{Q}^\top \mu_l/\tau)}$$

where $\bar{Q} = \frac{1}{H' \times W'} \sum_{i=1}^{H' \times W'} F_{s/t,i}$ represents the mean feature representation of a source or target image, obtained by averaging the feature vectors $F_{s/t,i}$ over all spatial locations ($i = 1, 2, \ldots, H' \times W'$). μ_k denotes the prototype vector for class k, τ is a temperature parameter, and K is the total number of classes.

This regularization term aims to minimize the negative log-likelihood of the mean feature representation $\bar{Q}$ being assigned to its corresponding class prototype μ_k. By maximizing the similarity between $\bar{Q}$ and μ_k while minimizing the similarity with other class prototypes, the regularization term encourages the feature representations to be globally diverse and well-separated.

4.4.1 Centroid-aware Pixel Contrast

Here, [30] introduce centroid-aware pixel contrast, namely SePiCo (ProtoCL).

rotoCL ($M = N = 1$). Naively operate K global category prototypes to establish one positive pair and $K - 1$ negative pairs. [30] consider this formulation as the prototype pixel contrast loss function:

$$\ell^{protocl} q = -\log \frac{e^{q^\top \mu^+/\tau}}{e^{q^\top \mu^+/\tau} + \sum\limits_{k \in \mathcal{K}^-} e^{q^\top \mu_k^-/\tau}}$$

where μ^+ is the positive prototype belonging to the same category as the specific query q and μ_k^- is the prototype of the k-th different category.

In summary, both global prototypes and local centroids serve as effective contrastive samples for learning pixel representations in the embedding space. They pull similar pixel representations closer while pushing dissimilar ones apart. The main distinction between these two approaches lies in the number of positive and negative pairs considered. Intuitively, if the quantity of contrastive pairs plays a significant role, it is reasonable to hypothesize that an infinite number of such pairs would contribute to the formation of a more robust and discriminative embedding space. This assumption will be further justified from a distributional standpoint.

4.4.2 Distribution-aware Pixel Contrast

[30] introduce a new contrastive loss that considers an infinite number of positive and negative pixel pairs for each pixel representation across source and target domains. Explicitly sampling M examples from the same latent class and N examples from other classes becomes computationally infeasible for large M and N due to GPU memory constraints. The complexity also grows exponentially with the number of semantic classes, making it impractical for tasks with many categories. Balanced sampling of positive and negative examples from each class is challenging, especially for imbalanced or long-tailed distributions. This condensed version captures the key points while using standard notation like [**?**] for references, and mathematical symbols like M and N. It omits some of the detailed explanations to make the passage more concise.

To address this issue, [30] take an infinity limit on M and N, where their effect is absorbed probabilistically. As M, N goes to infinity, it becomes the estimation of:

$$\begin{aligned}
\ell^\infty q &= \lim_{M \to \infty} \lim_{N \to \infty} \ell^{\text{cl}} q \\
&= \lim_{M \to \infty} -\frac{1}{M} \sum_{m=1}^{M} \log \frac{e^{q^\top q^+/\tau}}{e^{q^\top q^+/\tau} + \sum\limits_{k \in \mathcal{K}^-} \lim\limits_{N \to \infty} \frac{1}{N} \sum\limits_{n=1}^{N} e^{q^\top q_{k,n}^-/\tau}} \\
&= -\mathbb{E} q^+ \sim p(q^+) \log \frac{e^{q^\top q^+/\tau}}{e^{q^\top q^+/\tau} + \sum\limits_{k \in \mathcal{K}^-} \mathbb{E}_{q_k^- \sim p(q_k^-)} e^{q^\top q_k^-/\tau}}
\end{aligned}$$

Algorithm 1: SCALE algorithm

Input: Source images Is, labels Ys, target images It, parameters λ_{cl}, λ_{reg}, and maximum/warm-up iterations L/Lw

Initialize encoder Θ_e with ImageNet pretrained parameters and randomly initialize classifier Θ_c;
Initialize class statistics $\{\mu_k\}_{k=1}^{K}$ and $\{\Sigma_k\}_{k=1}^{K}$ to zeros;
Teachers init: $\Theta'_e \leftarrow \Theta_e$, $\Theta'_c \leftarrow \Theta_c$;
for $iter \leftarrow 0$ **to** L **do**
    Randomly sample a source image Is with Ys and a target image It;
    Obtain image features Fs and Ft;
    Update class means $\{\mu_k\}_{k=1}^{K}$ and covariance matrices $\{\Sigma_k\}_{k=1}^{K}$ via or memory bank with current centroids $\{\mu'_k\}_{k=1}^{K}$;
    if $iter > Lw$ **then**
        Train Θ_c using L_{ce}, L_{ssl}, L_{cl}, L_{reg};
    end
    else
        Train Θ_c using L_{ce}, L_{ssl};
    end
    Update teachers Θ'_c with Θ_c ;
end
return *Final classifier* Θ_c *weights*

where $p(q^+)$ is the positive semantic distribution with the same label as q and $p(q_k^-)$ is the k-th negative semantic distribution with a different label from each query q. The analytic form above is intractable, but it has a rigorous closed-form upper bound, which can be derived:

$$\begin{aligned}
\ell^{\infty} q &= -\mathbb{E}q^+ \log \frac{e^{q^\top q^+/\tau}}{e^{q^\top q^+/\tau} + \sum_{k\in\mathcal{K}^-} \mathbb{E}q_k^- e^{q^\top q_k^-/\tau}} \\
&\le \log \left[\mathbb{E}q^+ \left[e^{q^\top q^+/\tau} + \sum_{k\in\mathcal{K}^-} \mathbb{E}q_k^- e^{q^\top q_k^-/\tau} \right] \right] - q^\top \mathbb{E}q^+ \left[\frac{q^+}{\tau} \right] \\
&= \log \left[\mathbb{E}q^+ e^{q^\top q^+/\tau} + \sum_{k\in\mathcal{K}^-} \mathbb{E}q_k^- e^{q^\top q_k^-/\tau} \right] - q^\top \mathbb{E}q^+ \left[\frac{q^+}{\tau} \right] \\
&= \ell_q^{distcl}
\end{aligned}$$

where the inequality follows from Jensen's inequality on concave functions, i.e., $\mathbb{E}\log(X) \le \log\mathbb{E}(X)$. Thus, the distribution-aware pixel contrast loss, i.e., SePiCo (DistCL), is yielded to implicitly explore infinite samples.

Next, to facilitate our formulation, [30] further assume the feature distribution. For any random variable x that follows Gaussian distribution $x \sim \mathcal{N}(\mu, \Sigma)$, [] have the moment-generating function that satisfies:

$$\mathbb{E}\left[e^{a^\top x} \right] = e^{a^\top \mu + \frac{1}{2} a^\top \Sigma a}$$

where μ is the expectation of x, Σ is the covariance matrix of x. Under the assumption that $q^+ \sim \mathcal{N}(\mu^+, \Sigma^+)$ and $q_k^- \sim \mathcal{N}(\mu_k^-, \Sigma_k^-)$, where μ^+ and Σ^+ denote the mean and covariance matrix of the positive semantic distribution for q, and μ_k^- and Σ_k^- represent the statistics of the k-th negative distribution, the following equation for a given pixel representation q simplifies to:

$$\begin{aligned}
\ell^{distcl} q &= \log \left(e^{\frac{q^\top \mu^+}{\tau} + \frac{q^\top \Sigma^+ q}{2\tau^2}} + \sum_{k\in\mathcal{K}^-} e^{\frac{q^\top \mu_k^-}{\tau} + \frac{q^\top \Sigma_k^- q}{2\tau^2}} \right) - \frac{q^\top \mu^+}{\tau} \\
&= -\log \frac{e^{\frac{q^\top \mu^+}{\tau} + \frac{q^\top \Sigma^+ q}{2\tau^2}}}{e^{\frac{q^\top \mu^+}{\tau} + \frac{q^\top \Sigma^+ q}{2\tau^2}} + \sum_{k\in\mathcal{K}^-} e^{\frac{q^\top \mu_k^-}{\tau} + \frac{q^\top \Sigma_k^- q}{2\tau^2}}} + \frac{q^\top \Sigma^+ q}{2\tau^2}
\end{aligned}$$

As a result, the overall loss function for each pixel-wise representation reduces to a closed form whose gradients can be analytically computed. This formulation allows for the efficient optimization of the contrastive loss, as the gradients can be directly calculated without the need for approximation techniques or sampling strategies.

The derived closed-form expression for the distribution-aware contrastive loss enables the model to effectively learn pixel-level representations by considering the statistics of the positive and negative semantic distributions[30]. By incorporating the mean and covariance information of these distributions, the loss function captures the inherent structure and variability of the data, leading to more robust and discriminative pixel embeddings.

4.4.3 Training Procedure

Our proposed SCALE method aims to learn a discriminative embedding space, which is complementary to the self-training approach[19]. To stabilize training and yield discriminative features that promote the generalization ability of the model, we unify both SCALE and self-training into a one-stage, end-to-end pipeline. The overall training objective is formulated as:

$$\min_{\Theta_e, \Theta_c, \Theta_p} L_{ce} + L_{ssl} + \lambda_{cl} L_{cl} + \lambda_{reg} L_{reg}$$

where Θ_e, Θ_c, and Θ_p denote the parameters of the encoder, classifier, and projection head, respectively. L_{ce} represents the cross-entropy loss for supervised learning, L_{ssl} is the self-supervised learning loss, L_{cl} is the contrastive loss for semantic alignment, and L_{reg} is the regularization term for encouraging diverse and smooth feature representations. The constants λ_{cl} and λ_{reg} control the strength of the corresponding losses[17]. Initial tests suggest that using equal weights to combine L_{cl} and L_{reg} yields better results. For simplicity, both are set to 1.0.

By optimizing the above objective function, clusters of pixels belonging to the same category are pulled together in the feature space while simultaneously being pushed apart from other categories[26]. This process establishes a discriminative embedding space that minimizes the gap across domains and enhances intra-class compactness and inter-class separability within a unified framework. Consequently, this approach facilitates the generation of reliable pseudo labels, which in turn

CE loss	SSL loss	Dist loss	Reg loss	Avg Acc ResNet-50 (%)	Avg Acc ResNet-108 (%)	Avg Acc ResNet-150 (%)
✓	–	–	–	61.0±0.6	70.0±0.4	72.0±0.7
✓	–	–	–	60.0±1.1	70.5±1.3	73.5±0.5
✓	–	✓	–	66.0±1.2	75.0±1.1	78.0±0.5
✓	–	–	✓	61.5±0.7	71.5±1.1	75.5±0.6
✓	✓	✓	–	74.0±1.2	86.0±1.2	87.5±0.3
✓	–	✓	✓	70.0±0.8	81.0±0.7	84.5±0.3
✓	✓	✓	✓	75.5±0.4	84.5±0.5	85.0±0.5

Table 1: Choice of different loss and base classifier

Method	A>C(%)	A>P(%)	A>R(%)	C>P(%)	C>R(%)	P>R(%)	C>A(%)	P>A(%)	R>A(%)	P>C(%)	R>C(%)	R>P(%)	Avg(%)
None	48.11	62.94	72.50	62.18	64.88	73.10	54.76	51.09	66.30	45.36	51.39	78.26	60.91
CLIP	51.60	81.90	82.60	81.90	82.60	71.90	71.90	71.90	71.90	51.60	51.60	81.90	71.10
DAPrompt	54.10	84.30	84.80	83.70	85.90	84.80	74.40	83.70	75.20	54.60	54.70	83.80	75.33
AD-CLIP	55.40	85.21	85.62	85.81	86.20	85.41	76.10	76.70	76.81	56.10	56.11	85.50	75.91
SCALE (Ours)	**64.40**	**80.51**	**80.35**	**78.68**	**76.50**	**80.67**	**73.51**	**71.51**	**77.58**	**64.94**	**68.26**	**84.24**	**75.09**

Table 2: Comparison of SCALE with existing methods

benefits the self-training process. The integration of SCALE and self-training in a single-stage, end-to-end pipeline allows for the simultaneous optimization of both objectives, leading to more stable training and improved generalization performance.

5 Experiments

In this section, we conducted ablation experiments to test the impact on model performance when several different losses, included CE loss, SSL loss, Dist loss, and Reg loss, are introduced. And we compare the performance of our method with existing methods. We completed the full experimental procedure on the Office-Home dataset. More details will be explained as following.

5.1 Dataset

Office-Home is a benchmark dataset for domain adaptation which contains 4 domains where each domain consists of 65 categories. The four domains are: Art – artistic images in the form of sketches, paintings, ornamentation, etc.; Clipart – collection of clipart images; Product – images of objects without a background and Real-World – images of objects captured with a regular camera. It contains 15,500 images, with an average of around 70 images per class and a maximum of 99 images in a class.

5.2 Model Setting

Both the teacher and student networks choose ResNet-50 as the encoder and use a two-layer fully connected network as the classification header with Relu as the activation function and an output layer size of 65, and a two-layer fully connected network as the projection header with relu as the activation function and an output layer size of 128.

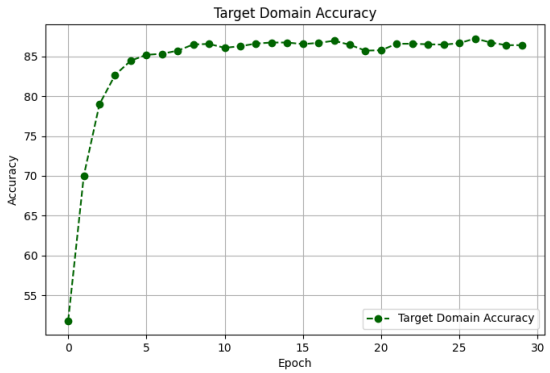

Figure 2: target accuracy

Regarding the loss function, we set an adaptive loss SSL loss for a threshold value of 0.8, and the loss is computed only when the confidence of the prediction result exceeds this value. We set the temperature coefficient of Reg loss to 0.1, and all the above parameters are the best parameters we have chosen after cross-validation.

5.3 Experiment Results

In this investigation, we delve into the intricacies of domain adaptation by leveraging each of the four distinct domains as a source domain, while scrutinizing the resultant classification outcomes on the target domain (comprising the remaining three domains). Our initial analysis compares the average classification accuracy attained by models trained solely on the source domain datasets, devoid of supplementary methods, yielding a modest 60.91% accuracy on the target domain.

During our rigorous testing phase, a noteworthy observation emerged: the direct migration from the Real World dataset to the Product dataset exhibited unparalleled performance, boasting an impressive accuracy of 78.26%. Similarly, the reciprocal transition from the Product dataset to the Real World

dataset yielded commendable results, underscoring the high degree of similarity between these two datasets. This empirical evidence corroborates our initial hypothesis regarding the dataset homogeneity.

Further experimentation involved benchmarking our proposed SCALE framework against several classical methods in the domain adaptivity realm, including CLIP, DAPrompt, and AD-CLIP. Notably, our comprehensive evaluation revealed that the SCALE framework nearly matches the performance of existing optimal methods across various scenarios.

Of particular significance are the tasks involving transitions from Art to Clipart, Product to Clipart, and Real World to Clipart datasets, where conventional methods faltered. Intriguingly, our SCALE framework demonstrated superior efficacy in these domains, outperforming all existing methodologies. This disparity in performance underscores the unique challenges posed by the Clipart dataset, suggesting its distinctiveness from the other three datasets. These findings collectively reinforce the efficacy and versatility of our proposed framework in tackling complex domain adaptation tasks.Detailed results are shown in Table 2

5.4 Ablation Study

In this section, we tested the classification accuracy on the target domain based on introducing only the cross-entropy loss CE loss computed from the source domain prediction results and the real labels, and then gradually introduced the self-training loss SSL loss, the distribution contrastive loss, Dist loss, and the regularisation loss Lreg loss to test the classification accuracy on the target domain, respectively. The results show that introducing only the self-training loss or regularisation loss does not significantly improve the classification results, and only when the self-training loss SSL loss, distribution contrast loss Dist loss are introduced at the same time, the model shows a significant performance improvement, and in particular, when the regularisation loss Lreg loss is introduced, not only can it bring about an accuracy improvement of 1%-1.5%, but also makes the model's training process is more stable and reduces the risk of overfitting.In addition to this, our previous test results were performed based on the ResNet-50 model as an encoder for images, and we tested the performance of SCALE when using a different model as an image encoder in our ablation experiments. The results of all ablation experiments are shown in Table 1.

6 Discussion

In this paper, we propose a novel adaptive framework for image classification called SCALE, which aims to achieve knowledge transfer in cross-domain image classification tasks through semantic contrast-aware learning. SCALE combines key steps such as semantic segmentation, contrastive learning, feature extraction and deep neural network image classification, and adopts a teacher-student architecture to further improve the performance of the model in the self-training process. We thus conducted extensive experiments on the Office-Home dataset. The experimental results show that our method achieves an average classification accuracy of 75.09%, which is only about 3% lower than the baseline method. At the same time, the model performance of SCALE far exceeds Baseline method, thus demonstrating the success of SCALE in domain transfer learning capabilities. While achieving excellent results, our approach still has some limitations, such as computation inefficiency. Currently, our model requires the entire source domain dataset to compute the distribution characteristics, which can be computationally expensive and memory-intensive. Future research should explore more efficient algorithms or techniques that can compute the distribution characteristics without loading the entire dataset into memory. In addition, there are several future research directions to consider. Investigating the transferability of learned representations across different domains would be valuable. This could involve exploring transfer learning techniques or domain adaptation methods to improve the model's ability to generalize to new and unseen domains.

References

[1] Ben-David, Shai, et al. "Analysis of representations for domain adaptation." In Advances in neural information processing systems 19, 2006.

[2] Zhang, Yuchen, et al. "Bridging theory and algorithm for domain adaptation." In International conference on machine learning, PMLR, 2019.

[3] Ganin, Yaroslav, et al. "Domain-adversarial training of neural networks." Journal of machine learning research 17.59 (2016): 1-35.

[4] Ganin, Yaroslav, and Victor Lempitsky. "Unsupervised domain adaptation by backpropagation." In International conference on machine learning, PMLR, 2015.

[5] Yu, Chaohui, et al. "Learning to match distributions for domain adaptation." arXiv preprint arXiv:2007.10791 (2020).

[6] Monteiro, Joao, et al. "Domain conditional predictors for domain adaptation." In NeurIPS 2020 Workshop on Preregistration in Machine Learning, PMLR, 2021.

[7] Lu, Wang, et al. "Fixed: Frustratingly easy domain generalization with mixup." In Conference on Parsimony and Learning, PMLR, 2024.

[8] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531 (2015).

[9] Adriana, Romero, et al. "Fitnets: Hints for thin deep nets." In Proc. ICLR 2.3, 2015: 1.

[10] Touvron, Hugo, et al. "Training data-efficient image transformers & distillation through attention." In International conference on machine learning, PMLR, 2021.

[11] Yang, Jing, et al. "Knowledge distillation via softmax regression representation learning." In International Conference on Learning Representations (ICLR), 2021.

[12] Zhao, Borui, et al. "Decoupled knowledge distillation." In Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, 2022.

[13] Wu, Haiyan, et al. "Self-supervised models are good teaching assistants for vision transformers." In International Conference on Machine Learning, PMLR, 2022.

[14] Bai, Yutong, et al. "Masked autoencoders enable efficient knowledge distillers." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[15] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016

[16] Targ, Sasha, Diogo Almeida, and Kevin Lyman. "Resnet in resnet: Generalizing residual architectures." arXiv preprint arXiv:1603.08029 (2016)

[17] Dollár, Piotr, Mannat Singh, and Ross Girshick. "Fast and accurate model scaling." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

[18] Lee, Jungkyu, et al. "Compounding the performance improvements of assembled techniques in a convolutional neural network." arXiv preprint arXiv:2001.06268 (2020).

[19] Touvron, Hugo, et al. "Fixing the train-test resolution discrepancy." Advances in neural information processing systems 32 (2019).

[20] Yuan, Li, et al. "Tokens-to-token vit: Training vision transformers from scratch on imagenet." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[21] Caron, Mathilde, et al. "Emerging properties in self-supervised vision transformers." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[22] Zou, Yang, et al. "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training." Proceedings of the European conference on computer vision (ECCV). 2018.

[23] Dong, Jiahua, et al. "Where and how to transfer: Knowledge aggregation-induced transferability perception for unsupervised domain adaptation." IEEE Transactions on Pattern Analysis and Machine Intelligence 46.3 (2021): 1664-1681.

[24] Corbiere, Charles, et al. "Confidence estimation via auxiliary models." IEEE Transactions on Pattern Analysis and Machine Intelligence 44.10 (2021): 6043-6055.

[25] Araslanov, Nikita, and Stefan Roth. "Self-supervised augmentation consistency for adapting semantic segmentation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

[26] Mei, Ke, et al. "Instance adaptive self-training for unsupervised domain adaptation." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16. Springer International Publishing, 2020.

[27] Kang, Guoliang, et al. "Pixel-level cycle association: A new perspective for domain adaptive semantic segmentation." Advances in neural information processing systems 33 (2020): 3569-3580.

[28] Wang, Qin, et al. "Domain adaptive semantic segmentation with self-supervised depth estimation." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[29] Wang, Wenguan, et al. "Exploring cross-image pixel contrast for semantic segmentation." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[30] Xie, Binhui, et al. "Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation." IEEE Transactions on Pattern Analysis and Machine Intelligence (2023).