Robust Semi-Supervised Learning for Self-learning Open-World Classes

Wenjuan Xi[1], Xin Song[2], Weili Guo[1,†], Yang Yang[1,3,†]

[1]Nanjing University of Science and Technology, China
[2]Baidu Talent Intelligence Center, China
[3]Hong Kong Polytechnic University, China
{xiwenjuan, wlguo, yyang}@njust.edu.cn, songxin06@baidu.com

Abstract—Existing semi-supervised learning (SSL) methods assume that labeled and unlabeled data share the same class space. However, in real-world applications, unlabeled data always contain classes not present in the labeled set, which may cause classification performance degradation of known classes. Therefore, open-world SSL approaches are researched to handle the presence of multiple unknown classes in the unlabeled data, which aims to accurately classify known classes while fine-grained distinguishing different unknown classes. To address this challenge, in this paper, we propose an open-world SSL method for Self-learning Open-world Classes (SSOC), which can explicitly self-learn multiple unknown classes. Specifically, SSOC first defines class center tokens for both known and unknown classes and autonomously learns token representations according to all samples with the cross-attention mechanism. To effectively discover novel classes, SSOC further designs a pairwise similarity loss in addition to the entropy loss, which can wisely exploit the information available in unlabeled data from instances' predictions and relationships. Extensive experiments demonstrate that SSOC outperforms the state-of-the-art baselines on multiple popular classification benchmarks. Specifically, on the ImageNet-100 dataset with a novel ratio of 90%, SSOC achieves a remarkable 16% improvement.

Index Terms—open-world semi-supervised learning, self-learning
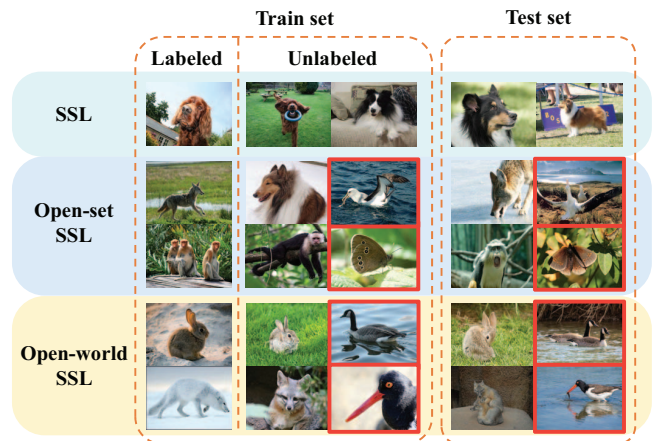
Fig. 1: SSL settings do not include the concept of novel classes. In open-set SSL, the unknown classes in the unlabeled dataset are treated as a single category, and exclusively focus on known class classification. In open-world SSL, multiple unknown classes can be distinguished. Unknown class samples are highlighted with a red border.

I. INTRODUCTION

With the development of machine learning, deep learning has achieved significant accomplishments in various domains such as vision, text, and speech [1], [2], [3], [4]. In the early stages, supervised learning relied on a large amount of expensive labeled data for model training. SSL emerged to reduce costs by effectively utilizing abundant unlabeled data [5], achieving performance comparable to supervised learning in scenarios with limited labeled data [6], [7]. However, almost all SSL methods are based on a default assumption: unlabeled data and labeled data are sampled from the same distribution, and there are no unseen classes for the model, as shown in the first row of Fig. 1. While this assumption has enabled their widespread application in closed-world scenarios [8], [9], it leads to significant performance degradation in real-world situations [10], [11], [12], [13]. For instance, in pathological image analysis, some lesion tissue slide images may come from unknown diseases, each with significant variations. This requires models that can distinguish unknown pathological image categories. In the field of network security, security experts use models for detecting malicious software, with the expectation that these models can differentiate new types of malicious software [14], [15]. Therefore, it is essential to investigate a more inclusive and versatile open-world approach.

Recently, researchers have proposed the concept of open-world SSL [16]. This framework allows for the presence of unknown classes in the unlabeled dataset, which are classes that do not appear in the labeled set, as shown in Fig. 1, third row. The objective is to classify both known and unknown classes simultaneously. Previously, open-set SSL methods have been proposed to address the real scenario [17], [18]. However, unlike the open-world SSL approach discussed in this paper, open-set SSL methods simply reject unknown class samples to prevent them from negatively affecting the classification performance of known classes in the unlabeled dataset. Their primary objective remains classifying the known classes, as depicted in the second row of Figure 1. Similarly, novel class discovery methods [19], [20], [21] also consider the open-world setting, but they assume that the unlabeled

---

† Corresponding authors.

dataset contains only unknown class samples and focus solely on clustering the unknown classes without considering the model's performance on known classes. Compared to these two approaches, open-world SSL poses a more challenging problem. The key to addressing this problem lies in enabling the model to learn multiple unknown classes better while ensuring the classification performance of known classes.

The problem of open-world SSL has attracted extensive attention, and several methods have been proposed to address this issue [16], [22]. These methods primarily approach the problem from the perspective of loss functions, employing optimization objectives that facilitate the learning of unknown classes and incorporating uncertainty mechanisms or adaptive thresholds to alleviate the issue of class imbalance during the learning process [23]. Although these methods have achieved remarkable performance in open-world SSL, they are limited to implicitly modeling multiple novel classes and partitioning the novel class space at the label level. This approach makes it difficult for the model to truly understand category concepts, leading to biases and typically exhibiting poor robustness when facing unknown classes. Therefore, there is still a lack of an open-world SSL method that can explicitly model category concepts, effectively discriminate between categories, and uncover hidden patterns and structures in the data.

To address this, we propose an open-world SSL method called Self-Supervised Open-World Class (SSOC) that aims to explicitly self-learn multiple unknown classes. Specifically, we initialize prototypes (class centers) representations for both known and unknown classes. Then, we iteratively learn the representation of class centers by combining a cross-attention mechanism with data features to achieve explicit modeling of class information. To assist the learning of multiple novel classes, we employ confident unlabeled samples to constrain the entropy loss and utilize pairwise similarity loss to extract information from unlabeled data, ensuring consistency at both the instance-level and prediction-level representations. The model architecture of SSOC not only ensures the classification performance of known classes but also focuses on capturing the differences and similarities among classes. By learning the feature representations of class prototypes, SSOC can perform classification operations with clear physical interpretations. This provides a new direction for exploring the application of interpretable learning methods in open-world SSL tasks.

In summary, our work has the following contributions:

- We propose an open-world SSL method for self-learning open-world classes. The method exploits a cross-attention mechanism to model class concepts explicitly and learn multiple unknown classes autonomously.
- We design a paired similarity loss to intelligently utilize information from unlabeled data for novel class discovery via instance prediction and relationship identification.
- We conducted experiments on CIFAR-10, CIFAR-100, and ImageNet-100 datasets with various data partitions to illustrate the effectiveness of SSOC. Our results showcase the remarkable robustness of SSOC in scenarios with limited labeled data and many novel classes.

II. RELATED WORK

Open-world SSL is closely related to SSL, open-set SSL, and novel class discovery (NCD). In this section, we summarize the similarities and differences among these research directions and investigate their development history.

A. Semi-Supervised Learning

The goal of SSL is to improve the learning performance of models by utilizing a large amount of unlabeled data along with a small amount of labeled data, thereby addressing the problem of high annotation costs. In recent years, significant progress has been made in deep SSL methods, which can be categorized into two main approaches: consistency regularization [24] and pseudo-labeling methods [25], [26]. Consistency regularization is based on the core idea that the model's predictions should remain consistent or similar under small perturbations of input data. It aims to enforce the model to produce consistent outputs for perturbed versions of the same input, promoting robustness and generalization. On the other hand, pseudo-labeling methods aim to augment the labeled dataset by using the model's predictions on unlabeled data as pseudo-labels. These pseudo-labels are treated as ground-truth labels for the unlabeled data, expanding the training dataset and enabling the model to learn from the additional unlabeled samples. Many methods combine both ideas. For example, MixMatch [27] applies random data augmentation to unlabeled samples and uses MixUp [28] pseudo-labeling to leverage information from unlabeled data. FixMatch [29] uses strong augmented images along with pseudo-labels generated from weakly augmented versions to learn consistency targets. However, SSL methods tend to misclassify unknown samples as known classes, which leads to a significant performance degradation in open-world scenarios. Consequently, traditional SSL methods are not suitable for effectively addressing open-world problems.

B. Open-World Semi-Supervised Learning

To address the emergence of novel classes, researchers have started studying the open-world problem. Initially, unknown classes were treated as a single category, leading to the proposal of the open-set SSL scenario. This scenario assumes that the unlabeled dataset may contain samples from unknown classes, and the goal is to reduce the negative impact of unknown classes on the learning of known classes and improve the model's robustness. In recent years, several open-set SSL methods have emerged, focusing on reducing the weight of unknown class samples [30] or selectively using unknown class samples [31], [32]. For example, DS^3L [30] decreases the weight of unknown class samples to reduce interference from unknown classes, while S2OSC [31] combines the sampling of novel class samples and model retraining in an adaptive manner, effectively incorporating semi-supervised learning. However, in practical applications, there may be multiple novel classes, and research has proposed adding post-processing steps for unknown class clustering to distinguish multiple novel classes after open-set SSL methods [16]. However, this

approach does not yield satisfactory results due to the lack of focus on unknown classes during training.

To address these issues, a recent work [16] introduces the open-world SSL setting, which assumes the presence of multiple unknown classes in the unlabeled dataset. The goal is to classify known classes while discovering multiple unknown classes. Compared to previous research, the open-world SSL scenario is closer to real-world situations but is still in its early stages. ORCA [16] is the first end-to-end deep learning framework proposed to tackle this problem. It establishes an uncertainty-adaptive margin mechanism to enhance the learning of unknown classes and achieves outstanding results. NACH [22] designs adaptive thresholds to balance the learning of known and unknown classes and proposes a novel classification loss function to aid the model in learning unknown classes, achieving better performance than ORCA.

C. Novel Class Discovery

The open-world SSL task's essence is discovering multiple unknown novel classes, closely related to NCD, which falls under weakly supervised learning [19], [20], [21], [33]. It involves training with both labeled and unlabeled datasets and assumes that both the unlabeled training and test sets comprise samples from unknown classes, with the goal of clustering these unknown classes. Initially, NCD methods typically employ a two-stage learning approach, where prior knowledge is learned on the labeled dataset, and then transfer learning-like methods are used to cluster the unknown classes in the unlabeled set. For instance, CILF [33] proposed a framework for incremental learning that acquires adaptive embeddings to tackle novel class detection, while DTC [34] proposed a deep transfer clustering method that leverages prior knowledge from known classes to enhance the representation of unknown classes and learns class prototypes to accomplish the clustering of unknown classes. However, NCD methods solely concentrate on the clustering performance of unknown classes while disregarding the classification task of known classes. Consequently, when both known and unknown classes are present in the test set, NCD methods struggle to achieve excellent overall performance.

III. PRELIMINARY

In the context of open-world SSL, the training set consists of a labeled dataset $\mathcal{D}^l = \{(x_i, y_i)\}_{i=1}^{\mathcal{M}}$ containing $\mathcal{M}$ labeled samples and an unlabeled dataset $\mathcal{D}^u = \{x_i\}_{i=1}^{\mathcal{N}}$ containing $\mathcal{N}$ unlabeled samples, where $x \in \mathbb{R}^D$ and D represents the image input dimension. We define $\mathcal{C}^l$ as the set of classes that appear in $\mathcal{D}^l$, and $\mathcal{C}^u$ as the set of classes that appear in $\mathcal{D}^u$. In the labeled dataset $\mathcal{D}^l$, $y \in \mathcal{C}^l = \{1, ..., |\mathcal{C}^l|\}$, while in the unlabeled dataset $\mathcal{D}^u$, x belongs to a certain class in $\mathcal{C}^u$. Assuming $\mathcal{C}^l \neq \mathcal{C}^u$ and $\mathcal{C}^l \cap \mathcal{C}^u \neq \emptyset$, we define $\mathcal{C}^s = \mathcal{C}^l \cap \mathcal{C}^u$ as the set of known classes and $\mathcal{C}^n = \mathcal{C}^u \setminus \mathcal{C}^l$ as the set of unknown classes. $S = |\mathcal{C}^s|$ and $N = |\mathcal{C}^n|$ represent the number of known and unknown classes, respectively.

In the previously mentioned methods related to open-world SSL, there is no concept of novel class in SSL, and it assumes $\mathcal{C}^l = \mathcal{C}^u$ by default, while NCD assumes that the unlabeled data does not contain any known classes, that is to say, $\mathcal{C}^l \cap \mathcal{C}^u = \emptyset$. Therefore, open-world SSL is inherently more challenging in nature.

IV. METHOD

In this section, we will first introduce the model architecture for self-supervised open-world classification. Then, we will formalize several learning objectives aimed at discovering multiple unknown classes. Finally, we will provide an overall algorithmic framework.

A. Self-Learning Open-World Classes

From the formalized problem definition, it can be understood that the key to the open-world SSL problem lies in effectively utilizing the unknown class samples in the unlabeled dataset. The main idea of SSOC is to explicitly self-learn open-world classes, regardless of whether they are known or unknown, which means we aim to learn representations of classes of the same dimensionality as image features. Obtaining class prototype representations alone is not a difficult problem; many previous unsupervised clustering methods can partition the data into several clusters and compute the center of each cluster. However, these methods do not utilize any labeled information and simply optimize the feature embedding model to achieve good clustering performance without incorporating class centers into the learning process. During the training process of open-world SSL, the data is input to the model in batches, and we expect the model to discover class information in each batch as much as possible and dynamically adjust the class centers. In backpropagation, explicit label information and the potential similarity relationships between samples are applied to optimize the class centers and the feature representations of images. To achieve this functionality, the core module of SSOC employs a cross-attention mechanism, which is the only parameterized network layer in SSOC apart from the feature extractor.

The cross-attention mechanism is a method that captures the correlation features between two sequences. It was first introduced in Transformer [35] and is used to fuse the input sequence from the decoder with the output sequence from the encoder, obtaining encoder information relevant to the current position of the decoder. The cross-attention mechanism has significantly contributed to sequence modeling and natural language processing tasks. For example, it has been used in image-text classification to merge multi-modal input sequences [36] and in machine translation to capture dependencies between distant positions in a sequence [37]. The internal structure of the cross-attention mechanism consists of three matrices: the query matrix W^Q, the key matrix W^K, and the value matrix W^V. The input sequences to the model can come from different modalities but must have the same dimensions. Using the query, key, and value matrices, the cross-attention mechanism calculates the correlation between sequences and fuses the value sequence relevant to the query sequence based on their weights. This mechanism allows
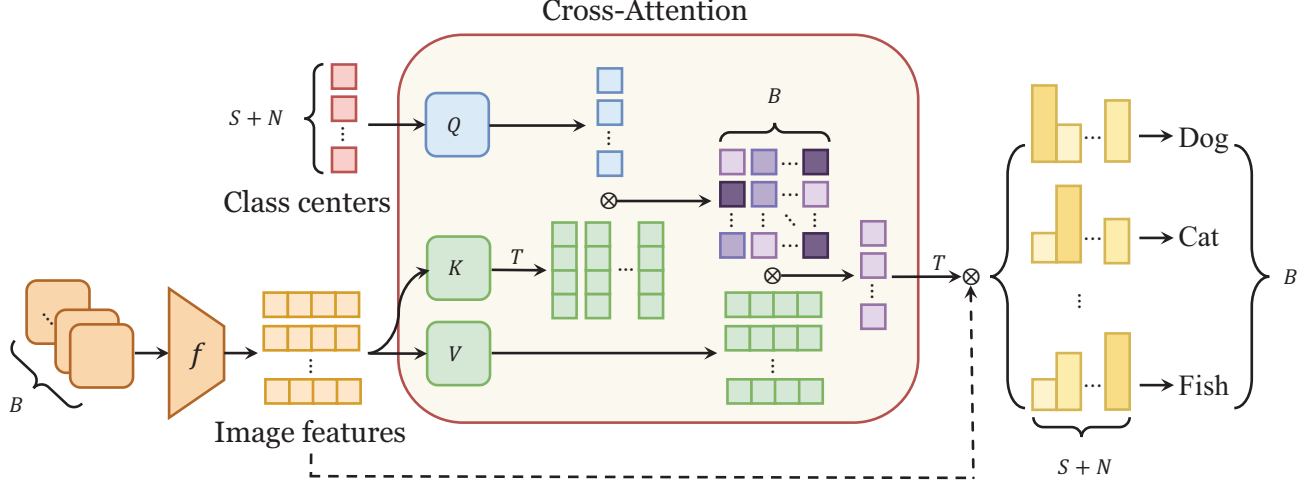
Fig. 2: SSOC framework. SSOC models the concept of explicit classes and utilizes a cross-attention mechanism to facilitate dynamic interaction between image features and class centers. Through iterative steps, SSOC achieves self-learning of class centers. During inference, SSOC predicts the class based on the similarity between image features and class centers.

the model to prioritize vital information in the sequences, enhancing task performance and capability.

In the context of this paper, we assume that $X = \{x_i | x_i \in \mathbb{R}^D\}_{i=1}^B$ represents the image data of the t-th batch, where B represents the batch size. By utilizing a pre-trained deep neural network $f_\theta : \mathbb{R}^D \to \mathbb{R}^d$, we obtain the embedding features $\mathcal{Z} = \{z_i = f_\theta(x_i) | x_i \in X, z_i \in \mathbb{R}^d\}$. Here, θ denotes the model parameters, and d represents the dimensionality of the embedding vectors. For ease of subsequent computations, SSOC sets the dimensions of W^Q, W^K, and W^V as $d \times d$. We denote the class centers obtained from the t-th batch as $\mathcal{A}_t = \{a_i\}_{i=1}^{S+N}$, particularly, $\mathcal{A}_0$ denotes the randomly initialized class center feature matrix, and a_i represents the feature vector of the i-th class. In the cross-attention mechanism, we treat class features as query inputs, and data features as key and value inputs, then do matrix multiplication with the corresponding parameter matrix, namely $Q = \mathcal{A}_t W^Q, K = \mathcal{Z}W^K, V = \mathcal{Z}W^V$. The employed cross-attention mechanism can be expressed as follows.

$$\Delta\mathcal{A} = \text{SoftMax}(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

The product of two vectors can usually represent the degree of similarity between them. In the equation above, QK^T represents the attention matrix of size $(S + N) \times B$. The element at the i-th row and j-th column of this matrix can be interpreted as the correlation between the i-th class center and the j-th sample in the batch. A higher value indicates that the sample is likelier to belong to that class. The attention scores are then passed through a SoftMax layer and weighted sum with the data features, resulting in the cross-attention matrix $\Delta\mathcal{A}$ of size $(S + N) \times d$. Each row of $\Delta\mathcal{A}$ represents the weighted sum of the i-th class center and the data features, indicating the contribution of samples more similar to the i-

th class in the batch. Conversely, samples less similar to the i-th class have smaller contributions. Therefore, $\Delta\mathcal{A}$ can be approximated as the class center feature vectors for the batch of samples. By computing the residual with the previous class centers, we can obtain the updated class centers $\mathcal{A}_{t+1}$.

$$\mathcal{A}_{t+1} = \mathcal{A}_t + \Delta\mathcal{A} \tag{2}$$

In this way, $\mathcal{A}_{t+1}$ retains the class information of most historical data while introducing the information of newly discovered classes in a residual manner. In the next iteration, it will be involved in computing the attention scores as class centers from the previous step. After several iterations, the class centers obtained from Equation 2 can represent the category concepts of all the data. Through this explicit learning approach, SSOC utilizes the cross-attention mechanism to achieve dynamic interaction between data features and class centers, enabling self-learning of categories. Compared to simply using clustering to obtain class centers, our method is less susceptible to the influence of data with extreme distribution deviations and better captures relevant features between samples and different classes.

In the preceding sections, we obtained the residual representation of class centers, denoted as $\Delta\mathcal{A}$. Subsequently, SSOC employs a distance-based approach, utilizing dot product similarity to measure the distance between a sample and a class center within the same feature space for interpretable classification. The distances are then normalized into a probability distribution using the activation function. Through our experiments, we observed the challenge of achieving a balanced learning process between known and unknown classes, often resulting in slow learning of unknown classes and premature overfitting to known classes. To address this issue, we adjusted the temperature parameter of the SoftMax function when calculating the predicted probabilities for labeled data, while

retaining the default value of 1 for unlabeled data. This adjustment involved utilizing a larger temperature parameter, leading to a flatter probability distribution, whereas a smaller value sharpens the output. By modifying the temperature parameter, our objective was to reduce the model's confidence in predicting known classes, thereby mitigating overfitting to known classes and enhancing the learning of unknown classes. Consequently, for a sample x_i, its probability distribution can be expressed as follows.

$$p_i = \begin{cases} \sigma(z_i \cdot \Delta \mathcal{A}^T / \epsilon) & x_i \in \mathcal{D}^l \\ \sigma(z_i \cdot \Delta \mathcal{A}^T) & \text{Others.} \end{cases} \quad (3)$$

where σ represents the SoftMax optimization operator, ϵ is the scaling hyperparameter, and p_{ij} represents the predicted probability of sample x_i belonging to class j. In the subsequent discussions, we employ $\hat{p}_i = \max_j p_i$ to denote the maximum confidence score and $\hat{y}_i = \arg\max p_i$ to represent the pseudo-label of the sample, which is then converted into one-hot form. Lastly, we illustrate the architecture of SSOC in Fig. 2

B. Optimize Objective

To assist SSOC in learning unknown classes in the open-world scenario, we designed a set of optimization objectives. Specifically, our loss function consists of three components: a cross-entropy loss that selects confident unlabeled data, a pairwise similarity loss that selects confident pairs of related samples, and a regularization term that prevents overfitting to known classes. The overall objective is as follows.

$$\mathcal{L} = \mathcal{L}_{CE} + \alpha \mathcal{L}_{BCE} + \beta \mathcal{L}_{RE} \quad (4)$$

where α and β are balancing hyper-parameters. Next, we will explain these three loss objectives in detail.

Cross-Entropy. Cross-entropy is a metric widely used to measure the difference between two probability distributions, commonly applied in classification tasks and probability estimation. For labeled data, SSOC maximizes the utilization of label information by minimizing the cross-entropy loss. The ground-truth labels of the labeled data $\mathcal{D}^l$ are represented as one-hot vectors Y^l, and the probability distribution P^l of the labeled data is calculated using Eq. 3. The supervised loss term can be expressed as follows.

$$\mathcal{L}_{CE}^l = -\frac{1}{\mathcal{M}} \sum_{i=1}^{\mathcal{M}} \sum_{j=1}^{S+N} Y_{ij}^l \log P_{ij}^l \quad (5)$$

For unlabeled data, they lack the necessary ground-truth labels, and direct computation of cross-entropy is impossible. To address this issue, many SSL methods employ pseudo-labels to compute the pseudo-supervised loss for unlabeled samples, helping the model extract information from unlabeled data. However, in many scenarios, noisy data in the dataset can result in incorrect or unreliable pseudo-labels, which can interfere with the model and degrade performance. To reduce noise interference and ensure the robustness of the model, we use a threshold to filter out high-confidence pseudo-labels for model training. Specifically, by calculating the probability

distribution P^u and pseudo-labels $\hat{Y}^u$ for the unlabeled sample set $\mathcal{D}^u$ using Eq. 3, we only consider unlabeled samples with maximum confidence scores higher than the threshold τ_1 for computing the cross-entropy loss.

$$\mathcal{L}_{CE}^u = -\frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \mathbb{I}(\hat{P}_i^u > \tau_1) \log \hat{P}_i^u \quad (6)$$

where $\mathbb{I}(\cdot)$ is the indicator function. In summary, the overall cross-entropy loss of SSOC is the weighted sum of the two components above.

$$\mathcal{L}_{CE} = \gamma \mathcal{L}_{CE}^l + \delta \mathcal{L}_{CE}^u \quad (7)$$

where γ and δ are balancing hyper-parameters.

In multiple experiments, we observed that during the initial training stages, the model has limited knowledge about unknown classes and exhibits higher uncertainty when predicting unlabeled data. As a result, the model initially selects a small number of unknown class samples, which is insufficient to develop a comprehensive understanding of unknown classes. However, the labeled data has precise and unambiguous ground-truth labels, and the supervision from $\mathcal{L}_{CE}^l$ is relatively strong. This can cause the model to be biased towards known classes and incorrectly classify unknown class samples as single known classes. To mitigate this issue and enhance the model's attention towards unknown classes, we introduce perturbations to the unlabeled data. For each unlabeled image sample, we generate two augmented images, r_1 and r_2, and concatenate their feature vectors with those of the labeled data. These concatenated feature vectors are collectively used to compute the class center increments in Eq. 1. When calculating the pseudo-supervised loss, we retain only the image data of r_1, treating the pseudo-labels of r_2 as the pseudo-labels of r_1. This approach enables the model to learn more invariant features of unknown classes.

Pairwise Similarity. Binary cross-entropy (BCE) loss is commonly used for binary classification tasks. To enable SSOC to learn better category features, we employ BCE loss to constrain the similarity between sample pairs. This idea has been successfully implemented in previous work such as NCD [38], [39]. In the embedding space, there are only two possible relationships between samples: same or different classes. The BCE loss aims to bring similar samples closer and push dissimilar samples apart. For labeled data, we directly use the ground-truth labels to determine whether they belong to the same class. For unlabeled data, we measure the similarity between sample pairs using the cosine similarity of their feature embeddings. To mitigate the negative impact of unreliable noise samples, we introduce a threshold τ_2 to filter out sample pairs with sufficient confidence.

$$\mathcal{L}_{BCE} = - \sum_{i=1}^{\mathcal{M}+\mathcal{N}} \sum_{j=1}^{\mathcal{M}+\mathcal{N}} \mathbb{I}(\min(\hat{P}_i, \hat{P}_j) > \tau_2)$$
$$\times [S_{ij} \log P_i^T P_j + (1 - S_{ij}) \log(1 - P_i^T P_j)] \quad (8)$$

$$S_{ij} = \begin{cases} \mathbb{I}(y_i = y_j) & x_i, x_j \in \mathcal{D}^l \\ \cos(\mathcal{Z}_i, \mathcal{Z}_j) & \text{Others.} \end{cases} \quad (9)$$

where $P = P^l \cup P^u$ represents the probability distribution of all data, and $\mathcal{Z} = \mathcal{Z}^l \cup \mathcal{Z}^u$ represents the features of all data.

During the minimization of the BCE loss, sample pairs with highly similar features (i.e., s_{ij} close to 1) have their probability distributions optimized to be more similar. Conversely, sample pairs with larger feature differences have their probability distributions optimized to be more distinct. Therefore, in our approach, the BCE loss aligns the prediction space with the embedding space, aiding the model in learning inter-class differences and intra-class similarities.

Maximum Entropy Regularization. In our experiments, we observed that during the initial training stage of SSOC, the cross-entropy loss dominated, resulting in the clustering of class centroids and impeding their separation. As a result, all unlabeled data might be mistakenly classified into a single category, which is undesirable. To promote a more uniform distribution of predicted classes, we introduced a maximum entropy regularization term to increase the uncertainty of the model's predictions. Maximum entropy regularization is essentially the empirical entropy, which estimates the uncertainty of the prior probability distribution based on the observed data frequencies. It can be expressed as follows.

$$\mathcal{L}_{RE} = \sum_{i=1}^{\mathcal{M}+\mathcal{N}} p_i \log p_i \quad (10)$$

In SSOC, we apply the above equation to all samples, maximizing the empirical entropy to make the model's predicted distribution more flexible and diverse, thus providing more opportunities for assigning unlabeled data to different classes. Experimental results demonstrate that the maximum entropy regularization term effectively enhances the model's robustness on unknown classes. In Algorithm 1, we provide a detailed description of the training process of SSOC.

V. EXPERIMENT

This section describes the experimental setup for SSOC, including datasets, comparison methods, implementation details, and evaluation metrics. We present experimental results on various data splits, validate loss function effectiveness through ablation study, and assess parameter impact.

A. Experimental Setup

Dataset. To validate the effectiveness of SSOC, we conducted experiments on three widely used computer vision benchmark datasets: CIFAR-10 [40], CIFAR-100 [40], and ImageNet-100 [41]. The CIFAR-10 and CIFAR-100 datasets consist of 60,000 images with a resolution of 32×32. Among them, 50,000 images are used for training and 10,000 for testing. CIFAR-10 consists of 10 classes, with approximately 6,000 images per class, while CIFAR-100 contains 100 classes, with around 600 images per class. ImageNet-100 is a subset of the ILSVRC2012 dataset, consisting of 100 classes selected

---

Algorithm 1 Robust Semi-Supervised Learning for Self-learning Open-World Classes

---

Input: Labeled datasets $\mathcal{D}^l = \{(x_i, y_i)\}_{i=1}^{\mathcal{M}}$, unlabeled dataset $\mathcal{D}^u = \{x_i\}_{i=1}^{\mathcal{N}}$, the seen classes number S, the novel classes number N, the number of iterations for updating class centers t, the current class centers $\mathcal{A}_t = \{a_i\}_{i=1}^{S+N}$, a pre-trained backbone f_θ, the initialized cross-attention layers g_φ, the number of the training epochs T, the mini-batch number B.

Output: A fine-tuned model f_θ and the trained cross-attention layers g_φ.

1: **for** epoch from 1 to T **do**
2:     **for** batch from 1 to B **do**
3:         $X^l, Y^l \leftarrow$ Sample mini-batch from $\mathcal{D}^l$
4:         $X_1^u, X_2^u \leftarrow$ Sample and augment mini-batch from $\mathcal{D}^u$
5:         $\mathcal{Z}^l, \mathcal{Z}_1^u, \mathcal{Z}_2^u \leftarrow f_\theta(X^l), f_\theta(X_1^u), f_\theta(X_2^u)$
6:         $\Delta\mathcal{A} \leftarrow g_\varphi(\mathcal{Z}^l \cup \mathcal{Z}_1^u \cup \mathcal{Z}_2^u, \mathcal{A}_t)$ with (Eq. 1)
7:         Compute P^l, P_1^u, P_2^u with (Eq. 3)
8:         $\hat{Y}_2^u \leftarrow \arg\max(P_2^u)$
9:         $\mathcal{Z}, P, Y \leftarrow \mathcal{Z}^l \cup \mathcal{Z}_1^u, P^l \cup P_1^u, Y^l \cup \hat{Y}_2^u$
10:       Compute $\mathcal{L}_{CE}$ with (Eq. 7)
11:       Compute $\mathcal{L}_{BCE}$ with (Eq. 8)
12:       Compute $\mathcal{L}_{RE}$ with (Eq. 10)
13:       Compute updated class centers $\mathcal{A}_{t+1}$ with (Eq. 2)
14:       $f_\theta, g_\varphi \leftarrow$ Optimize with (Eq. 4)
15:     **end for**
16: **end for**

---

from the original 1,000 classes [42]. For consistent comparison with other studies, we used the same 100 classes as ORCA and NACH. In all datasets, we applied random cropping and rotation as data augmentation techniques to augment the dataset. In the main experiments, we treated the first 50% of classes in each dataset as known classes and the remaining as unknown classes. We used 50% of the labeled data from the known classes as labeled data, while the rest, along with the data from the unknown classes, were treated as unlabeled data. Additionally, we presented results for different label ratios and novel ratios. Random data splits were used in all experiments to ensure the generalizability of the results.

Comparison method. We compare SSOC with SSL, open-set SSL, NCD, and existing open-world SSL methods. SSL and open-set SSL methods are limited to classifying known classes. To extend them to open-world SSL scenarios, K-means clustering of unknown classes is performed, allowing evaluation of their performance on novel classes. FixMatch [29] is chosen as the representative SSL method, and DS³L [30] and CGDL [32] as representative open-set SSL methods. Since SSL methods do not include the concept of novel classes, samples with low confidence in SoftMax outputs are treated as unknown classes. For NCD-based methods, DTC [34] and RankStats [39] are selected for comparison. Since they can only cluster unknown classes and lack performance

on known classes, we use the Hungarian algorithm to perform maximum weighted matching between the known classes in the labeled data and the clustered classes [43], [16], [22], and evaluate the results separately for these known classes. For open-world SSL, we compare the reported performances of ORCA and NACH, the two existing methods in this research area, to the best of our knowledge.

Implementation Details. For the CIFAR-10 dataset, we employ ResNet34 [44] as the backbone network and use two Adam optimizers to optimize the backbone network and the cross-attention layers separately. We fine-tune the backbone network with a lower learning rate of 1e-4, while the cross-attention layers are trained with a slightly higher learning rate of 5e-3, focusing on learning category information. Both optimizers have momentum parameters set to $(0.9, 0.99)$. We train the models with a batch size of 128 for 200 epochs. For the CIFAR-100 dataset, we use ResNet18 [44] as the backbone network. The backbone network and the cross-attention layers are trained with a learning rate of 1e-4. The batch size is 512, and the models are trained for 500 epochs. For the ImageNet-100 dataset, we adopt ResNet50 [44] as the backbone network. The backbone network and the cross-attention layers are trained with learning rates of 1e-5 and 3e-4, respectively. The models are trained with a batch size of 100 for 200 epochs. In all experiments, we employ an early stopping strategy and dynamically adjust the learning rate using the cosine annealing method. The experiments on the CIFAR dataset are conducted on 8 V100 GPUs, while the experiments on ImageNet-100 are performed using 4 NVIDIA 3090 GPUs.

We utilize a pre-trained ResNet model to extract high-quality image features. In our experiments, we extract the initial embedding vectors $\mathcal{Z}^l \cup \mathcal{Z}^u$ of all the data using the pre-trained backbone before training. Subsequently, we apply the K-means++ algorithm [45] to these vectors for unsupervised clustering, resulting in the initialization of class center representations $\mathcal{A}_0$. This initialization, which incorporates prior knowledge, facilitates the learning process of the model. Furthermore, in practical scenarios, it might be infeasible to obtain samples for unknown classes, we can resort to random initialization of prototypes for the unknown classes, while employing the clustering method mentioned earlier to obtain prototypes for the known classes. The code for SSOC can be found at https://github.com/njustkmg/SSOC.

Evaluation Metrics. We adopt the evaluation approach used in [16], [22] and report the accuracy of SSOC on known classes, unknown classes, and all classes. Additionally, in the ablation experiments, we report the normalized mutual information (NMI) on the novel classes. It is worth noting that since the learned concepts of novel classes by the model are unordered, before computing the accuracy on unknown classes and all classes, we utilize the Hungarian maximum weighted matching algorithm [43] for label alignment to obtain the optimal matching between the clustering labels of unknown classes and the ground-truth labels.

B. Main Results

Comparison of main results. In Tab. (I), we display SSOC's classification accuracies alongside other methods on CIFAR-10, CIFAR-100, and ImageNet-100 with 50% label ratio and 50% novel ratio. SSOC consistently outperforms SSL, open-set SSL, and NCD methods in open-world SSL scenarios, achieving superior accuracy for known, unknown, and all classes across all datasets. Compared to the best-performing method, RankStats, we achieved improvements of 30% and 42.4% on all classes of CIFAR-100 and ImageNet-100, respectively. Furthermore, SSOC outperforms the two open-world SSL methods overall. Compared to NACH, we achieved a 3.8% improvement in accuracy for known classes in CIFAR-10, and significantly improved the performance for unknown classes in the challenging ImageNet-100 dataset, with accuracy improvements of 2.9% and 3.1%, respectively. Our experimental results demonstrate that SSOC effectively addresses the open-world SSL problem.

Change the label ratio. To demonstrate the effectiveness of SSOC in scenarios with a limited amount of labeled data, we fixed the novel ratio at 50% and compared the performance of ORCA, NACH, and SSOC with label ratios of 10% and 30%. Tab. II presents the accuracy for all classes, including results from the ORCA and NACH papers. It can be observed that as the labeled data decreases, the performance of all three methods declines. Notably, on CIFAR-10 and ImageNet-100, SSOC only incurs performance drops of 1.89% and 7.22%, respectively, when the label ratio decreases from 50% to 10%, while NACH experiences declines of 3.2% and 12.79%, respectively. Additionally, on the ImageNet-100 dataset, when the label ratios are 10% and 30%, SSOC achieves an overall accuracy that is 8.66% and 6.09% higher than NACH, respectively. Though our CIFAR-100 result with a 30% label ratio is slightly inferior to NACH, it remains 6.95% above ORCA. Overall, SSOC demonstrates strong robustness and performs well in situations with limited labeled data.

Change the novel ratio. By fixing the label ratio at 50%, we explored the impact of varying novel ratios on open-world SSL methods. Tab. III shows the overall class accuracy of ORCA, NACH, and SSOC at novel ratios of 10%, 30%, 70% and 90%. Since the ORCA and NACH papers lack these experimental data, we reproduced their open-source code. Observing the data, it is evident that as the number of unknown classes increases, the performance of all three methods decreases, but SSOC exhibits a smaller performance drop than the other two methods. Notably, SSOC achieves impressive results in scenarios with a higher novel ratio. When the novel ratio is 90%, SSOC achieves accuracy improvements of 10.61-22.3% compared to NACH on the three datasets, and a 16.01% accuracy improvement compared to ORCA on the ImageNet-100 dataset with a novel ratio of 90%. Furthermore, on the CIFAR dataset with a novel ratio of 10%, SSOC did not surpass NACH, indicating that NACH focuses more on classifying known classes, while SSOC places greater emphasis on discovering unknown classes and still achieves

664

TABLE I: Accuracy on CIFAR-10, CIFAR-100, and ImageNet-100, with 50% label ratio and 50% novel ratio. The best results are highlighted in bold.

Method Classes	Dataset	CIFAR-10			CIFAR-100			ImageNet-100		
	Method	Seen	Novel	All	Seen	Novel	All	Seen	Novel	All
SSL	FixMatch	71.5	50.4	49.5	39.6	23.5	20.3	65.8	36.7	34.9
Open-set SSL	DS^{3}L	77.6	45.3	40.2	55.1	23.7	24.0	71.2	32.5	30.8
	CGDL	72.3	44.6	39.7	49.3	22.5	23.5	67.3	33.8	31.9
NCD	DTC	53.9	39.5	38.3	31.3	22.9	18.3	25.6	20.8	21.3
	RankStats	86.6	81.0	82.9	36.4	28.4	23.1	47.3	28.7	40.3
Open-world SSL	ORCA	88.2	90.4	89.7	66.9	43.0	48.1	89.1	72.1	77.8
	NACH	89.5	92.2	91.3	68.7	47.0	52.1	91.0	75.5	79.6
	SSOC	**93.3**	**92.6**	**92.8**	**69.0**	**48.0**	**53.1**	**91.4**	**78.4**	**82.7**

TABLE II: Accuracy on CIFAR-10, CIFAR-100, and ImageNet-100 when changing the label ratio, with 50% novel ratio. The best results are highlighted in bold.

Dataset	CIFAR-10			CIFAR-100			ImageNet-100		
Label Ratio	ORCA	NACH	SSOC	ORCA	NACH	SSOC	ORCA	NACH	SSOC
10%	84.10	88.10	**90.90**	38.60	43.38	**43.82**	69.70	66.81	**75.47**
30%	87.70	90.21	**92.24**	43.71	**51.06**	50.66	70.79	70.74	**76.83**
50%	89.70	91.30	**92.79**	48.10	52.10	**53.13**	77.80	79.60	**82.69**

TABLE III: Accuracy on CIFAR-10, CIFAR-100, and ImageNet-100 when changing the novel ratio, with 50% label ratio. The best results are highlighted in bold.

Dataset	CIFAR-10			CIFAR-100			ImageNet-100		
Novel Ratio	ORCA	NACH	SSOC	ORCA	NACH	SSOC	ORCA	NACH	SSOC
10%	91.20	**95.84**	95.14	61.35	**66.77**	64.29	90.36	89.34	**91.58**
30%	91.08	93.08	**95.13**	52.17	54.12	**56.39**	86.09	80.03	**90.49**
50%	89.70	91.30	**92.79**	48.10	52.10	**53.13**	77.80	79.60	**82.69**
70%	81.52	86.44	**87.38**	38.12	43.93	**44.78**	57.57	53.88	**73.58**
90%	56.63	59.19	**75.52**	28.29	26.79	**37.40**	41.10	43.56	**65.86**

excellent results even with a small number of known classes.

Through these three experiments, we have demonstrated the effectiveness of SSOC and its outstanding robustness and generalization performance. SSOC outperforms ORCA and NACH in handling scenarios with limited labeled data or many novel classes, making it applicable in a wider range of scenarios and having significant practical implications.

C. Ablation Study

To validate the effectiveness of different loss functions, we conducted ablation experiments on ImageNet-100 with a label ratio and novel ratio of 50%. Tab. IV reports the accuracy on known classes, unknown classes and the overall accuracy, as well as the NMI on unknown classes. In the first three rows, we removed the cross-entropy loss, pairwise similarity loss, and maximum entropy regularization term, respectively, and used the remaining loss as the objective function. The experimental results show that $\mathcal{L}_{CE}$ plays a crucial role in SSOC. The model relies on the supervised

TABLE IV: Ablation study on ImageNet-100 dataset, we report the accuracy and NMI on unknown classes.

Approach	Seen	Novel	Novel(NMI)	All
w/o $\mathcal{L}_{CE}$	2.06	2.03	3.67	1.35
w/o $\mathcal{L}_{BCE}$	90.60	75.51	75.29	80.33
w/o $\mathcal{L}_{RE}$	81.48	37.77	63.88	47.00
SSOC	91.39	78.37	78.32	82.69

loss from labeled data to obtain essential ground-truth class information and learns about unknown classes through pseudo-supervised loss. Furthermore, in the experiment where $\mathcal{L}_{RE}$ was removed, the performance on novel classes significantly deteriorated, confirming the beneficial effect of maximum entropy regularization on learning novel classes.

D. Parametric Analysis

Effect of threshold selection. To analyze the impact of thresholds on the experimental results, we conducted experiments

(*a*) Selected pseudo-labels for unseen classes (different τ_1)

(*b*) Accuracy of pseudo-labels for unseen classes (different τ_1)

(*c*) Accuracy of pseudo-labels for unseen classes (different τ_2)
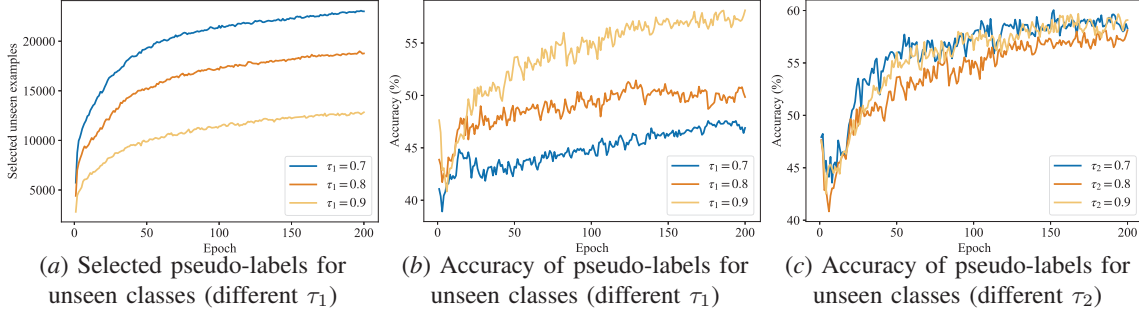
Fig. 3: Analysis of threshold effect on the CIFAR-100 dataset. Fig. 3(a) and Fig. 3(b) illustrate the number of selected pseudo-labels and the corresponding accuracy for different values of τ_1. Fig. 3(c) presents the accuracy of the selected pseudo-labels for different values of τ_2.

(*a*) Accuracy on seen classes

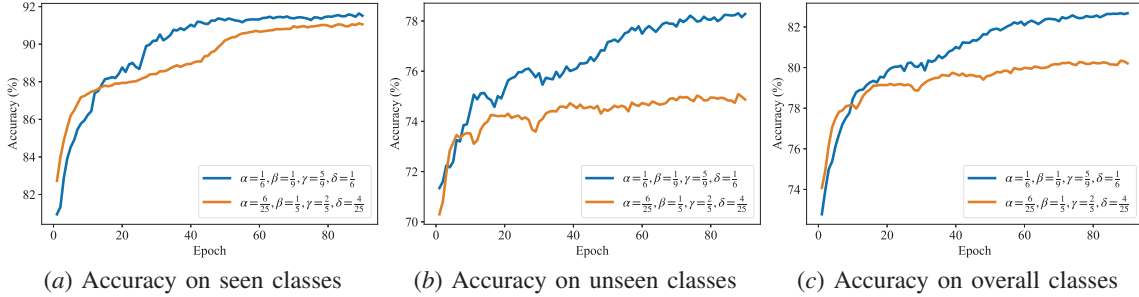(*b*) Accuracy on unseen classes

(*c*) Accuracy on overall classes

Fig. 4: Analysis of the effect of loss balance hyper-parameters on the ImageNet-100 dataset. We select two sets of parameter configurations, and in Fig. 4(a), Fig. 4(b), and Fig. 4(c), we present the accuracy on seen classes, unseen classes, and overall classes, respectively.

on the CIFAR-100 dataset with different values of τ_1 and τ_2, with a label ratio and novel ratio of 50%. In Fig. 3 (a) and (b), we varied the value of τ_1 and presented the number of unlabeled samples selected by the threshold during each epoch of training in (a), and the pseudo-label accuracy of selected unknown class samples in (b). It can be observed that a lower τ_1 is insufficient to filter out falsely labeled samples, which can interfere with the model's classification ability. On the other hand, a larger τ_1 excessively eliminates unlabeled data, failing to provide sufficient unknown class samples for the model to learn, resulting in poor performance on unknown classes. In (c), we provide the pseudo-label accuracy of selected unknown class samples for different τ_2 values. In all CIFAR-100 experiments, we set τ_1 to 0.6 and τ_2 to 0.8.

Effect of the balance hyper-parameter on the loss. To investigate the impact of different loss weights on the results, we present the accuracies of SSOC on the ImageNet-100 dataset for two sets of loss weights, in Fig. 4. It can be observed that elevating the weight of loss terms associated with unknown classes seems to constrain the model's capacity to learn about unknown classes, which subsequently impacts the overall performance. This suggests that the loss terms $\mathcal{L}^u_{CE}$, $\mathcal{L}_{BCE}$, and $\mathcal{L}_{RE}$ are closely related to the learning of unknown classes, and simply emphasizing the learning of unknown classes may hurt overall classification. We can assume an extreme case: when the supervised loss $\mathcal{L}^l_{CE}$ is negligible, SSOC degenerates into a clustering algorithm on unlabeled data,

without any information about ground-truth classes. When performing label matching using the Hungarian maximum weight algorithm, incorrect matches between predicted labels and true labels increase, resulting in suboptimal results for unknown classes and overall classification. Therefore, we need to find an optimal set of loss balance weights.

VI. CONCLUSION

In this work, we propose SSOC to address the open-world SSL problem. SSOC autonomously learns the categories in the open world using a cross-attention mechanism and leverages pairwise similarity loss to extract information from unlabeled data, discovering novel classes through instance prediction and relationships. We demonstrate the effectiveness of SSOC on three benchmark computer vision datasets, where it outperforms state-of-the-art baseline methods. Moreover, SSOC exhibits excellent robustness when facing challenges such as limited labeled data and many novel classes.

REFERENCES

[1] K. Aggarwal, M. M. Mijwil, A.-H. Al-Mistarehi, S. Alomari, M. Gök, A. M. Z. Alaabdin, S. H. Abdulrhman *et al.*, "Has the future started? the current growth of artificial intelligence, machine learning, and deep learning," *Iraqi Journal for Computer Science and Mathematics*, vol. 3, pp. 115–123, 2022.

[2] Y. Matsuo, Y. LeCun, M. Sahani, D. Precup, D. Silver, M. Sugiyama, E. Uchibe, and J. Morimoto, "Deep learning, reinforcement learning, and world models," *Neural Networks*, vol. 152, pp. 267–275, 2022.

[3] S. Pouyanfar, S. Sadiq, Y. Yan, H. Tian, Y. Tao, M. E. P. Reyes, M. Shyu, S. Chen, and S. S. Iyengar, "A survey on deep learning: Algorithms, techniques, and applications," *ACM Computing Surveys (CSUR)*, vol. 51, pp. 92:1–92:36, 2019.

[4] S. Dargan, M. Kumar, M. R. Ayyagari, and G. Kumar, "A survey of deep learning and its applications: a new paradigm to machine learning," *Archives of Computational Methods in Engineering*, vol. 27, pp. 1071–1092, 2020.

[5] Y. Yang, Z. Fu, D. Zhan, Z. Liu, and Y. Jiang, "Semi-supervised multi-modal multi-instance multi-label deep network with optimal transport," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, pp. 696–709, 2021.

[6] Y. Yang, D. Zhou, D. Zhan, H. Xiong, Y. Jiang, and J. Yang, "Cost-effective incremental deep model: Matching model capacity with the least sampling," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, pp. 3575–3588, 2023.

[7] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine learning*, vol. 109, pp. 373–440, 2020.

[8] Y. Yang, D. Zhan, Y. Wu, Z. Liu, H. Xiong, and Y. Jiang, "Semi-supervised multi-modal clustering and classification with incomplete modalities," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, pp. 682–695, 2021.

[9] Y. Yang, H. Wei, H. Zhu, D. Yu, H. Xiong, and J. Yang, "Exploiting cross-modal prediction and relation consistency for semi-supervised image captioning," *IEEE Transactions on Cybernetics*, 2021.

[10] M. Pezeshki, L. Fan, P. Brakel, A. C. Courville, and Y. Bengio, "Deconstructing the ladder network architecture," in *ICML*, New York, NY, 2016, pp. 2368–2376.

[11] M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," in *NeurIPS*, Barcelona, Spain, 2016, pp. 1163–1171.

[12] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *ICLR*, Toulon, France, 2017.

[13] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *NeurIPS*, Long Beach, CA, 2017, pp. 1195–1204.

[14] D. Zhou, Y. Yang, and D. Zhan, "Learning to classify with incremental new class," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, pp. 2429–2443, 2022.

[15] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," in *NeurIPS*, Quebec, Canada, 2015, pp. 3546–3554.

[16] K. Cao, M. Brbic, and J. Leskovec, "Open-world semi-supervised learning," in *ICLR*, Virtual Event, 2022.

[17] C. Geng, S.-j. Huang, and S. Chen, "Recent advances in open set recognition: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, pp. 3614–3631, 2020.

[18] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, pp. 1757–1772, 2013.

[19] C. Troisemaine, V. Lemaire, S. Gosselin, A. Reiffers-Masson, J. Flocon-Cholet, and S. Vaton, "Novel class discovery: an introduction and key concepts," *arXiv preprint arXiv:2302.12028*, 2023.

[20] P. Nodet, V. Lemaire, A. Bondu, A. Cornuéjols, and A. Ouorou, "From weakly supervised learning to biquality learning: an introduction," in *IJCNN*, Shenzhen, China, 2021, pp. 1–10.

[21] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National science review*, vol. 5, pp. 44–53, 2018.

[22] L. Guo, Y. Zhang, Z. Wu, J. Shao, and Y. Li, "Robust semi-supervised learning when not all classes have labels," in *NeurIPS*, New Orleans, LA, 2022, pp. 3305–3317.

[23] Y. Chen, X. Zhu, W. Li, and S. Gong, "Semi-supervised learning under class distribution mismatch," in *AAAI*, New York, NY, 2020, pp. 3569–3576.

[24] D. Lee, S. Kim, I. Kim, Y. Cheon, M. Cho, and W. Han, "Contrastive regularization for semi-supervised learning," in *CVPR*, New Orleans, LA, 2022, pp. 3910–3919.

[25] X. Yang, Z. Song, I. King, and Z. Xu, "A survey on deep semi-supervised learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, pp. 8934–8954, 2023.

[26] A. Oliver, A. Odena, C. Raffel, E. D. Cubuk, and I. J. Goodfellow, "Realistic evaluation of deep semi-supervised learning algorithms," in *NeurIPS*, S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., Montréal, Canada, 2018, pp. 3239–3250.

[27] D. Berthelot, N. Carlini, I. J. Goodfellow, N. Papernot, A. Oliver, and C. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *NeurIPS*, BC, Canada, 2019, pp. 5050–5060.

[28] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *ICLR*, BC, Canada, 2018.

[29] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. Raffel, E. D. Cubuk, A. Kurakin, and C. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *NeurIPS*, Virtual Event, 2020, pp. 596–608.

[30] L. Guo, Z. Zhang, Y. Jiang, Y. Li, and Z. Zhou, "Safe deep semi-supervised learning for unseen-class unlabeled data," in *ICML*, Virtual Event, 2020, pp. 3897–3906.

[31] Y. Yang, H. Wei, Z. Sun, G. Li, Y. Zhou, H. Xiong, and J. Yang, "S2OSC: A holistic semi-supervised approach for open set classification," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 16, pp. 34:1–34:27, 2022.

[32] X. Sun, Z. Yang, C. Zhang, K. V. Ling, and G. Peng, "Conditional gaussian distribution learning for open set recognition," in *CVPR*, Seattle, WA, 2020, pp. 13 477–13 486.

[33] Y. Yang, Z. Sun, H. Zhu, Y. Fu, Y. Zhou, H. Xiong, and J. Yang, "Learning adaptive embedding considering incremental class," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, pp. 2736–2749, 2023.

[34] K. Han, A. Vedaldi, and A. Zisserman, "Learning to discover novel visual categories via deep transfer clustering," in *ICCV*, Seoul, Korea (South), 2019, pp. 8400–8408.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, Long Beach, CA, 2017, pp. 5998–6008.

[36] A. Jaegle, S. Borgeaud, J. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, E. Shelhamer, O. J. Hénaff, M. M. Botvinick, A. Zisserman, O. Vinyals, and J. Carreira, "Perceiver IO: A general architecture for structured inputs & outputs," in *ICLR*, Virtual Event, 2022.

[37] M. Gheini, X. Ren, and J. May, "Cross-attention is all you need: Adapting pretrained transformers for machine translation," in *EMNLP*, Punta Cana, Dominican Republic, 2021, pp. 1754–1765.

[38] Z. Zhong, E. Fini, S. Roy, Z. Luo, E. Ricci, and N. Sebe, "Neighborhood contrastive learning for novel class discovery," in *CVPR*, Virtual Event, 2021, pp. 10 867–10 875.

[39] K. Han, S. Rebuffi, S. Ehrhardt, A. Vedaldi, and A. Zisserman, "Automatically discovering and learning new visual categories with ranking statistics," in *ICLR*, Addis Ababa, Ethiopia, 2020.

[40] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009. [Online]. Available: https://www.cs.toronto.edu/~kriz/cifar.html

[41] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[42] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, M. Proesmans, and L. Van Gool, "Scan: Learning to classify images without labels," in *ECCV*, Glasgow, UK, 2020, pp. 268–285.

[43] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, pp. 83–97, 1955.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, Las Vegas, NV, 2016, pp. 770–778.

[45] B. Bahmani, B. Moseley, A. Vattani, R. Kumar, and S. Vassilvitskii, "Scalable k-means++," *Proceedings of the VLDB Endowment*, vol. 5, pp. 622–633, 2012.