Unknown-Aware Graph Regularization for Robust Semi-supervised Learning from Uncurated Data

Heejo Kong[1], **Suneung Kim**[2], **Ho-Joong Kim**[2], **Seong-Whan Lee**[2]

[1]Department of Brain and Cognitive Engineering, Korea University, Seoul, South Korea
[2]Department of Artificial Intelligence, Korea University, Seoul, South Korea
hj_kong@korea.ac.kr, se_kim@korea.ac.kr, hojoong_kim@korea.ac.kr, sw.lee@korea.ac.kr

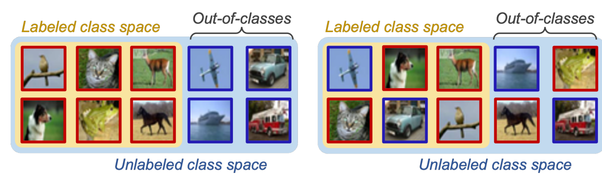
Abstract

Recent advances in semi-supervised learning (SSL) have relied on the optimistic assumption that labeled and unlabeled data share the same class distribution. However, this assumption is often violated in real-world scenarios, where unlabeled data may contain out-of-class samples. SSL with such uncurated unlabeled data leads training models to be corrupted. In this paper, we propose a robust SSL method for learning from uncurated real-world data within the context of open-set semi-supervised learning (OSSL). Unlike previous works that rely on feature similarity distance, our method exploits uncertainty in logits. By leveraging task-dependent predictions of logits, our method is capable of robust learning even in the presence of highly correlated outliers. Our key contribution is to present an unknown-aware graph regularization (UAG), a novel technique that enhances the performance of uncertainty-based OSSL frameworks. The technique addresses not only the conflict between training objectives for inliers and outliers but also the limitation of applying the same training rule for all outlier classes, which are existed on previous uncertainty-based approaches. Extensive experiments demonstrate that UAG surpasses state-of-the-art OSSL methods by a large margin across various protocols. Codes are available at https://github.com/heejokong/UAGreg.

Introduction

Recent advances in deep supervised learning have been driven by the availability of large-scale annotated datasets. However, constructing such training data is labor-intensive and time-consuming due to the labeling process. As a remedy, significant efforts have been dedicated to the field of semi-supervised learning (SSL) (Lee et al. 2013; Xie et al. 2020; Sohn et al. 2020; Li, Xiong, and Hoi 2021; Zheng et al. 2022). They have provided effective solutions to leverage abundant unlabeled data with only a fraction of manual annotations, and shown their promising performances.

All the positive results observed in SSL are typically based on the optimistic assumption that both labeled and unlabeled data are drawn from the identical class distribution. However, in practical scenarios, the unlabeled dataset often includes out-of-class data, *i.e.*, outlier, which easily violates this assumption. This uncurated unlabeled data severely degrade the performance of SSL (Oliver et al. 2018). Hence, it



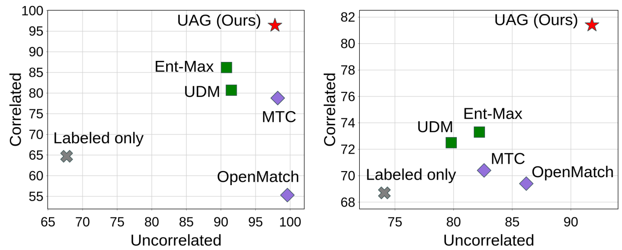
(a) Uncorrelated outliers (b) Correlated outliers

Figure 1: An illustration of the example for uncorrelated and correlated outliers. The classes marked with red and blue box share the same superclass, animal and transportation, respectively. In an uncorrelated setting, out-of-class data do not share a same superclass with labeled classes, whereas in a correlated setting, they share the same superclass space.

is desirable that the training models not only classify samples from known categories, *i.e.*, inliers, but also identify samples from novel classes as outliers. This task is known as open-set semi-supervised learning (OSSL) (Yu et al. 2020). While OSSL is more realistic and practical for various applications, it has been rarely considered in previous literature.

In OSSL problem, the challenge lies in the absence of supervision for distinguishing unknown from known samples. Recent notable approaches (Yu et al. 2020; Saito, Kim, and Saenko 2021) have tackled this problem by utilizing similarity distances in feature space. Based on intra- or inter-class distances of known categories, they aim to identify unknown samples, which deviate significantly from in-distribution (ID) data, and consider the samples as outliers. Although the techniques have substantially improved the performance of OSSL, they are still quite limited for general use. As discussed in (Saito, Kim, and Saenko 2021), similarity-based measures can fail to detect highly correlated outliers that exhibit similar visual characteristics to ID data. It is evident that this limitation is more pronounced when unknown classes share the same superclass space with inliers, defined as correlated outliers in Fig. 1. In a scenario where the scale of unlabeled data is larger, the correlated outliers are more natural, but it has never been considered in prior studies.

As an alternative to the feature similarity, some OSSL works (He et al. 2022; Sun and Wang 2022) exploit uncertainty in logit space to resolve the open-set decision prob-



(a) CIFAR-10 (b) CIFAR-100

Figure 2: AUROC of five OSSL algorithms trained on CIFAR datasets with 100 labels per class. The results are reported for two outlier settings, whether uncorrelated (x-axis) or correlated (y-axis). Methods marked with square (□) use the uncertainty in logits; methods with diamond ($\diamond$) use the similarity in features. The proposed method UAG (marked with $\star$) is based on uncertainty, outperforming the existing methods by a large margin.

lem. In contrast to the embedding features, the logits effectively activate task-specific information related to high-level semantic attributes (Wang et al. 2022), making uncertainty-based measures robust for highly correlated outliers. However, as shown in Fig. 2, the performance of uncertainty-based methods is still not competitive in practice.

We assert that the under-performing of uncertainty-based methods mainly originates from two aspects: *i)* Existing methods focus on maximizing the entropy of outliers to improve the discriminative ability for unknown contexts, while they aim to minimize the entropy of inliers for known contexts simultaneously. Despite the inherent conflict between the learning objectives, they solely rely on a single classifier. *ii)* Although the outliers are composed of multiple novel classes, previous works train the model by assigning them into one generic class, unknown. This approach leads to the convergence of all outliers into a single representative space, making it difficult to cope with a diverse range of outliers.

In this paper, we propose an unknown-aware graph regularization (UAG) as a novel method to address the drawbacks of uncertainty-based frameworks. To tackle the first aspect, we adopt a multi-head structure. The heads are trained on labeled data in the same way, while they aim to train inliers and outliers for unlabeled data independently. This approach effectively alleviates the conflict that arises when a single classifier head simultaneously learns entropy minimization and maximization. To mitigate the second aspect, we apply contrastive graph regularization. Based on batch-wise similarities for known and unknown contexts estimated from the multi-head, pseudo-label and pseudo-outlier graphs are constructed, and they are integrated into an unknown-aware pseudo graph. By employing contrastive learning on the integrated graph as a target, the model is trained to have similar embeddings for samples with similar both known and unknown contexts. This approach allows the outliers to have multiple clusters in the feature space, in contrast to the previous methods.

We conduct extensive experiments on CIFAR-10/100 (Krizhevsky, Hinton et al. 2009) and ImageNet-30 (Hendrycks et al. 2019) datasets, following the previous benchmarks. The results demonstrate that our method successfully addresses the limitations of uncertainty-based approaches and significantly surpasses the performance of previous state-of-the-art methods. Notably, our approach exhibits superior performance, even in scenarios where the outliers are highly correlated with the ID data.

Related Works

Semi-supervised learning. A standard SSL assumes that all training and testing data share identical class distribution, regardless of whether they are labeled or not, and aims to classify unlabeled examples into known classes. The mainstream of SSL can be broadly categorized into entropy minimization (Lee et al. 2013; Grandvalet and Bengio 2004; Cascante-Bonilla et al. 2021; Maeng et al. 2013), consistency regularization (Tarvainen and Valpola 2017; Sajjadi, Javanmardi, and Tasdizen 2016; Laine and Aila 2016; Miyato et al. 2018; Xie et al. 2020; Nam et al. 2020), and holistic methods (Berthelot et al. 2019b,a; Sohn et al. 2020; Kong et al. 2023). Recently, several studies (Li, Xiong, and Hoi 2021; Zheng et al. 2022) have attempted to combine consistency regularization with contrastive learning (Chen et al. 2020a; Khosla et al. 2020) by exploiting the instance-level similarity relationships as a target. Note that most outstanding works (Sohn et al. 2020; Li, Xiong, and Hoi 2021; Zheng et al. 2022) for SSL are based on self-training. Hence, when not all classes have labels in the training data, these methods will train outlier examples from unlabeled data as the known categories, and this learning causes the return of the corrupted SSL model (Oliver et al. 2018).

Open-set semi-supervised learning. A OSSL problem considers a more practical scenario using uncurated unlabeled data, where the training data often contains out-of-class instances. Early works focused solely on preventing the degradation of closed-set accuracy by adaptively assigning the weights of examples expected to be out-of-class data (Guo et al. 2020) or completely filtering them (Chen et al. 2020b). However, these methods do not have any objective for separating outliers from in-distribution data. Representative OSSL works (Yu et al. 2020; Huang et al. 2021; Saito, Kim, and Saenko 2021) adopted the similarity distance in feature space for addressing the limitation. In the concept of intra-class (Yu et al. 2020) or inter-class (Saito, Kim, and Saenko 2021) distance for known categories, they attempted to distinguish the outliers from known ones. As another alternative, some works exploited uncertainty scores of predictions in logit space using confidence (Sun and Wang 2022) and energy function (He et al. 2022; Liu et al. 2020). However, as discussed in this paper, the similarity-based methods are difficult to cope with highly correlated outliers, and uncertainty-based methods do not show competitive performance compared to (Yu et al. 2020; Saito, Kim, and Saenko 2021). In contrast, the proposed UAG exhibits relatively consistent performance improvement across all configurations, irrespective of the correlation between inlier and outlier classes.

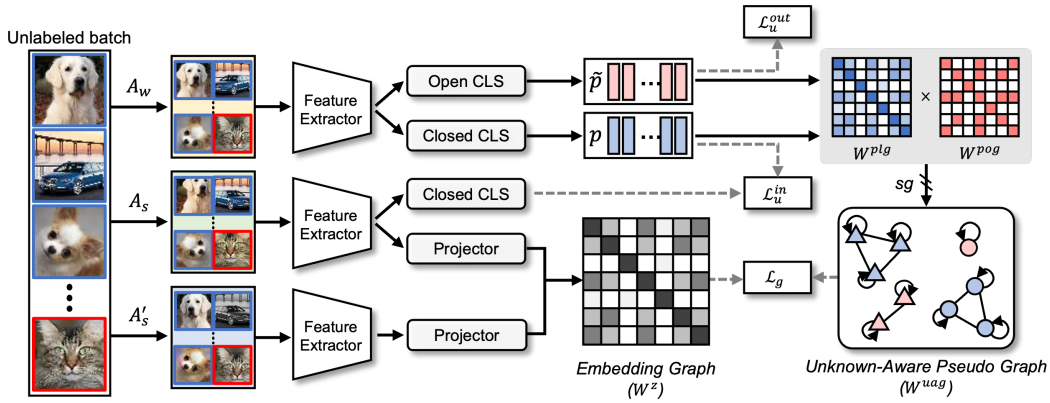


Figure 3: Framework of the proposed UAG. Weakly augmented views are used to generate model predictions for closed-set (p) and open-set ($\tilde{p}$) contexts. Based on $\tilde{p}$, pseudo-inliers and -outliers are assigned. p are used as targets for training of pseudo-inliers (Eq. 5), while $\tilde{p}$ is used to train pseudo-outliers (Eq. 6). Pseudo-label (W^{plg}) and pseudo-outlier (W^{pog}) graphs are constructed to measure similarity between samples for known and unknown contexts, respectively. Both graphs are integrated into an unknown-aware pseudo graph (W^{uag}) to train an embedding graph (Eq. 10).

Our Approach

Problem definition. Given a batch of labeled set $\mathcal{X} = \{(x_b, y_b)\}_{b=1}^{B}$, where y_b is the corresponding label, and a batch of unlabeled set $\mathcal{U} = \{(u_b)\}_{b=1}^{\mu B}$, where μ determines the relative size of $\mathcal{X}$ and $\mathcal{U}$, the objective of SSL is to learn a classification model by effectively leveraging both labeled and unlabeled data. Unlike standard SSL, we believe that the unlabeled data is more likely to be uncurated, where the unlabeled set may contain out-of-class data unseen in the labeled set. Hence, this work contributes to solving an OSSL problem. We aim to train a model not only to classify known K-way categories, *i.e.*, inlier, but also to identify novel classes, *i.e.*, outliers, from the known ones.

Approach overview. Our proposed model consists of a shared encoder $f(\cdot)$ with three heads: the closed-set classifier $h(\cdot)$, the open-set classifier $\tilde{h}(\cdot)$, and the projection head $g(\cdot)$. At test time, the closed-set classifier first predicts the K-way label for closed-set classification. Then, the open-set classifier measures the uncertainty of each prediction for open-set recognition. The projection head is deployed only for the training phase. A key technical novelty comes from the choice of a multi-head structure for enhancing the uncertainty-based approach as well as training them with contrastive graph regularization.

Uncertainty-based Outlier Detection

We exploit the uncertainty of logits to construct an outlier detector. Due to the mapping process into class-dependent space, the logits focus on activating only high-level semantic attributions for the target task (Wang et al. 2022). It effectively suppresses the interference of task-irrelevant correlation. Our work adopts a maximum softmax probability (MSP) (Hendrycks and Gimpel 2016) as an uncertainty score, and considers the samples with the low scores as outliers. Specifically, the score is derived from the composite function of encoder $f(\cdot)$ and the open-set classifier $\tilde{h}(\cdot)$ as follows:

$$s(x;T) = \max_i \frac{\exp\left([\tilde{h} \circ f(x)]_i / T\right)}{\sum_{j=1}^{K} \exp\left([\tilde{h} \circ f(x)]_j / T\right)}, \tag{1}$$

where T is a temperature parameter and set to larger than 1, for further enlarging the score gap between in- and out-of-distribution data, as discussed in (Liang, Li, and Srikant 2017). $[\tilde{h} \circ f(x)]_i$ indicates the logits of the i-th class.

To detect inliers and outliers with the estimated score $s(x;T)$, we utilize the thresholding, *i.e.*, the sample x is determined as inlier if $s(x;T) > \tau^{in}$ or outlier if $s(x;T) < \tau^{out}$. The thresholds are adaptively decided by employing a two-component Gaussian Mixture Model (GMM). For each training epoch t, GMM is fit on the predicted scores for the unlabeled set $\mathcal{U}$ with the Expectation-Maximization algorithm. The expectations of two gaussian distributions are assigned τ_t^{in} and τ_t^{out}, respectively. Instead of using the current thresholds directly, we employ EMA thresholds which are updated by averaging the consecutive ones with a momentum parameter m. Their initial values, $\hat{\tau}_0^{in}$ and $\hat{\tau}_0^{out}$, are set to 1. The EMA update is applied as a warm-up process to stabilize the thresholds at the early stage of the training as:

$$\hat{\tau}_t^{in} \leftarrow m\hat{\tau}_{t-1}^{in} + (1-m)\tau_t^{in}, \tag{2}$$

$$\hat{\tau}_t^{out} \leftarrow m\hat{\tau}_{t-1}^{out} + (1-m)\tau_t^{out}. \tag{3}$$

Unknown-Aware Graph Regularization

As discussed in introduction, this work aims to enhance the uncertainty-based OSSL framework by addressing two aspects: *i)* Conflicts between objectives for learning inliers and outliers, *ii)* The convergence of all outliers into a single representative space, even though the outliers may consist of various novel classes. Inspired by the above drawbacks, we propose a couple of learning approaches as follows, and an overview of the proposed framework is presented in Fig. 3.

	Corr. C10	Uncorr. C10
False positive ratio (FPR)	25.76	7.45
False negative ratio (FNR)	48.99	26.28

Table 1: Results of the models trained solely on labeled data for CIFAR-10 (100 labeled data per class). FPR/FNR (%) indicate the proportions of samples with erroneous predictions, even though they are confidently predicted as outliers and inliers. Note that the outlier type, whether uncorrelated (Uncorr.) or correlated (Corr.), is shown in each column.

Exclusive multi-head training. To address the first aspect, we adopt a multi-head structure that exclusively learns inliers and outliers. Note that existing methods (He et al. 2022; Sun and Wang 2022) only depend on a single classifier which trains to maximize the entropy of pseudo-outliers and minimize the entropy of pseudo-inliers. However, this strategy allows the model to learn a significant amount of erroneous predictions, as shown in Table 1. These false predictions act as noise that conflicts with our intended objective, as they are learned to minimize entropy even when outlier entropy should be maximized, and vice versa. As a result, the training models result in suboptimal naturally.

Whereas, in our method, two classifier heads learn the objectives for inliers and outliers independently. Specifically, they share the same feature extractor $f(\cdot)$, and train the labeled data as the cross-entropy loss $H(\cdot,\cdot)$ between ground-truth labels y_b and predictions:

$$\mathcal{L}_s = \frac{1}{B}\sum_{b=1}^{B}\left(\mathrm{H}(y_b, p(A(x_b))) + \mathrm{H}(y_b, \tilde{p}(A(x_b)))\right), \quad (4)$$

where $A(\cdot)$ refers to weak augmentation. p and $\tilde{p}$ represent the softmax probabilities predicted by closed-set classifier $h \circ f$ and open-set classifier $\tilde{h} \circ f$, respectively.

For training the closed-set classifier h, the samples predicted as inliers, *i.e.*, pseudo-inliers, are leveraged only. We adopt FixMatch (Sohn et al. 2020) as our learning objective due to its simplicity yet effectiveness. It is defined as the cross-entropy between pseudo-labels $\hat{y}_b$ and predictions:

$$\mathcal{L}_u^{in} = \frac{1}{\mu B}\sum_{b=1}^{\mu B}\mathbb{1}\left(s_b \geq \hat{\tau}_t^{in}\right)\cdot \mathrm{H}\left(\hat{y}_b, p(\hat{A}(u_b))\right), \quad (5)$$

where $\hat{A}(\cdot)$ refers to strong augmentation. $s_b = s(u_b; T)$ represents MSP score (Eq. 1) for b-th unlabeled example.

The open-set classifier $\tilde{h}$ learns only the examples assigned as outliers from unlabeled data. Following the previous work (Sun and Wang 2022), we directly maximize the mean entropy for the outliers, in order to separate them from the inliers in the task-dependent space. Specifically, we optimize the following objective:

$$\mathcal{L}_u^{out} = -\frac{1}{\mu B}\sum_{b=1}^{\mu B}\mathbb{1}\left(s_b < \hat{\tau}_t^{out}\right)\cdot \mathrm{H}\left(\tilde{p}(A(u_b))\right). \quad (6)$$

By contrast with Eq. 5, since outliers are not assigned any labels, only weak augmentation is utilized for stability.

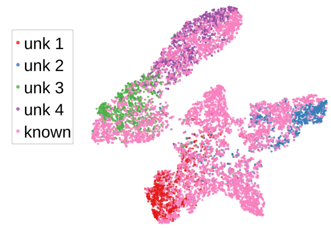


Figure 4: t-SNE visualization of embeddings obtained from a model trained with FixMatch. Pink points denote inliers, while the other colored points represent unknown classes.

Contrastive graph regularization. Existing works have adopted only the entropy maximization for entire outliers, despite the fact that they comprise multiple novel classes with diverse semantic information. When considering updates only for the weights $\omega \in \mathbb{R}^{k\times d}$ of the last classifier in this learning process, it can be expressed as follows. For simplicity, let's momentarily exclude the influence of bias.

$$\arg\min_{\omega\in\Omega} \mathbb{E}_{(x,y)\sim D}[\mathrm{H}(y, f(x,\theta);\omega)]. \quad (7)$$

By minimizing the cross-entropy loss for inliers, $\mathrm{H}(y, f(x,\theta); w) = -y\log\sigma\left(f(x,\theta)\cdot\omega^T\right)$, the weights converge to points in the embedding space that maximize the similarity with samples corresponding to the k-th label. In theoretical terms, the weights function as representative points for the embedding features $z = f(x,\theta)$ of each class.

$$\arg\max_{\omega\in\Omega} \mathbb{E}_{(u)\sim D}[\mathrm{H}(f(u,\theta);\omega)]. \quad (8)$$

From this perspective, maximizing the entropy of outliers aims to minimize the similarity between their embedding features z and the k-way representative points ω. Hence, the intermediate features of outlier examples converge to a single cluster orthogonal to the weights ω. Such training inhibits the model's ability to encode outlier data precisely, degrading the performance of outlier detection.

Interestingly, as shown in Fig. 4, we found that conventional SSL have low discriminative ability between outliers and inliers, while high discriminative power among outlier classes in embedding space. This suggests that semantic information about the known context can benefit learning about out-of-classes. Based on this empirical rationale, we derive an unknown-aware graph regularization that allows the outliers to form multiple clusters in their embeddings.

Specifically, following a previous CoMatch (Li, Xiong, and Hoi 2021), we build a pseudo-label graph $W^{plg} \in \mathbb{R}^{\mu B\times\mu B}$ by leveraging the batch-wise predictions of a closed-set classifier. The graph W^{plg} is a similarity matrix of the closed-set predictions, and assigns only examples with similar predictions for a known context to the same cluster:

$$W_{bj}^{plg} = \begin{cases} 1 & \text{if } b = j \\ p_b\cdot p_j & \text{if } b\neq j \text{ and } p_b\cdot p_j \geq \tau_g \\ 0 & \text{otherwise} \end{cases}, \quad (9)$$

where τ_g denotes a pre-defined threshold. p_b and p_j represent the softmax probabilities predicted by $h \circ f$ for u_b and u_j, respectively. Each sample is connected to itself with the strongest edge of value 1 as a self-loop, while the samples with similarity less than τ_g are not connected.

Unlike CoMatch, we utilize the pseudo-outlier graph $W^{pog} \in \mathbb{R}^{\mu B \times \mu B}$, which is information about the unknown context, in the process of forming the target graph. The graph W^{pog} is built by employing the batch-wise predictions of an open-set classifier. Note that $\eta_b = \mathbb{1}(s_b < \hat{\tau}_t^{out})$ is an outlier indicator, and returns 1 for values predicted as outliers and 0 for the others. The graph W^{pog} assigns only samples with the same prediction for the unknown context to the same cluster by connecting the edges with the same predictions of the outlier indicator to the value of 1 as:

$$W_{bj}^{pog} = \begin{cases} 1 & \text{if } \eta_b = \eta_j \\ 0 & \text{otherwise} \end{cases}. \tag{10}$$

An unknown-aware pseudo graph is obtained by matrix multiplication of two pseudo graphs: $W^{uag} = W^{plg} \cdot W^{pog}$. That is, only samples with the same predictions in unknown context as well as the similar predictions for known context are assigned to the same cluster in the graph. The graph W^{uag} serves as a target to train embedding graph.

To construct the embedding graph, we first obtain a pair of images with different augmentations $\hat{A}(\cdot)$ and $A'_s(\cdot)$ applied to each unlabeled sample. The embeddings are extracted from a composite function of encoder f and projection head g: $z_b = g \circ f(\hat{A}(u_b))$. The embedding graph W^z is derived as the batch-wise similarity of the two projected embeddings, where t_e is a scalar temperature parameter:

$$W_{bj}^{z} = \begin{cases} \exp(z_b \cdot z'_b / t_e) & \text{if } b = j \\ \exp(z_b \cdot z_j / t_e) & \text{otherwise} \end{cases}. \tag{11}$$

Subsequently, the graphs are normalized with $\hat{W}_{bj} = W_{bj} / \sum_j W_{bj}$, so that each row of the similarity matrix sums to 1. The graph contrastive loss is derived to minimize the cross-entropy between the two normalized graphs as:

$$\mathcal{L}_g = \frac{1}{\mu B} \sum_{b=1}^{\mu B} \mathrm{H}\left(\hat{W}_{bj}^{uag}, \hat{W}_{bj}^{z}\right). \tag{12}$$

By minimizing the loss $\mathcal{L}_g$, the embedding graph is enforced to have the same structure as W^{uag}. This learning encourages the model to have similar embeddings for samples with similar predictions in both closed-set and open-set recognition. In other words, outlier samples are positioned within a single space that minimizes their similarity with the known context, while samples with similar semantic information belong to the same cluster.

Overall objectives. The overall objective of our proposed framework is the weighted sum of the supervised loss $\mathcal{L}_s$, the unsupervised loss $\mathcal{L}_u$, and the contrastive loss $\mathcal{L}_g$. We summarize the unsupervised loss as the sum of $\mathcal{L}_u^{in}$ and $\mathcal{L}_u^{out}$. Hence, the overall loss can be written as follows:

$$\mathcal{L} = \mathcal{L}_s + \lambda_u \mathcal{L}_u + \lambda_g \mathcal{L}_g, \tag{13}$$

where λ_u and λ_g are the hyperparameters to control the weight of each objective.

Experiments

Experimental Settings

Setup. We follow the default setting of the previous work (Saito, Kim, and Saenko 2021): CIFAR-10/100 (Krizhevsky, Hinton et al. 2009) setup on WRN-28 (Zagoruyko and Komodakis 2016) for small-scale experiments; ImageNet (Deng et al. 2009) and Semi-iNature-2021 (Su and Maji 2021) setup on ResNet-18 (He et al. 2016) for large-scale experiments. Note that we use an identical set of hyperparameters for whole experiments except for projection head g, where the output dimension is scaled on each dataset. For an outlier detector, T and m are set to 1.5 and 0.9, respectively. The settings for graph contrastive loss are fixed across all experiments for simplicity, *e.g.*, a graph threshold τ_g, a scalar temperature t_e. More information regarding implementation details can be found in Appendix.

Compared methods. We compare the proposed UAG with representative SSL, similarity-based OSSL and uncertainty-based OSSL methods. For the SSL baselines, we employ FixMatch (Sohn et al. 2020) and CoMatch (Li, Xiong, and Hoi 2021). Since UAG is incrementally applied to these works (Sohn et al. 2020; Li, Xiong, and Hoi 2021), a comparison with them effectively reveals the strength of our proposed approach. As the similarity-based OSSL baselines, we employ MTC (Yu et al. 2020) and OpenMatch (Saito, Kim, and Saenko 2021) using the author's implementations. Referring to the papers for the uncertainty-based methods (He et al. 2022; Sun and Wang 2022), we implement two objectives, which are entropy maximization (Ent-Max) (Sun and Wang 2022) and uniform distribution matching (UDM) (He et al. 2022), and adopt the objectives in ablation studies to demonstrate the effectiveness of our proposed UAG. In addition, the most recent OSSL work, OSP (Wang et al. 2023), is also compared.

Evaluation metric. We evaluate our approach with two common metrics: Error rates for closed-set accuracy, and AUROC for open-set recognition. We reported the results averaged over three runs and their standard deviations.

Experimental Results

CIFAR-10/100. To assess the CIFAR datasets, we consider two setups: (a) uncorrelated, where unknown classes have no superclass relation with known ones, and (b) correlated, where known and unknown classes share superclass space. For CIFAR-10 (CIFAR-100), we divide classes into 6 known (60 known) and 4 unknown (40 unknown) classes. All samples, except for the labeled data, are designated as unlabeled data. Tables 2 and 3 show error rates and AUROC results, respectively. Our method consistently outperforms OSSL and SSL baselines in error rates. Notably, UAG shows significant AUROC improvement, contrasting with minimal gains from similarity-based OSSL baselines in the correlated setting, indicating UAG's robustness across outlier correlations.

ImageNet. We evaluate the performance of our proposed approach on ImageNet, a more challenging and complex dataset. Due to computational constraints, we utilize a subset called ImageNet-30 (Tack et al. 2020), which consists of 30 classes, for training instead of the complete ImageNet.

Dataset	CIFAR-10				CIFAR-100				ImageNet-30
	Uncorr.		Corr.		Uncorr.		Corr.		
No. of labeled	50	100	50	100	50	100	50	100	10%
Labeled Only	34.3±1.2	29.4±0.8	30.9±1.3	25.8±0.7	38.9±0.8	31.7±0.6	38.1±0.7	30.7±0.4	20.2±1.2
FixMatch	16.8±1.1	10.7±0.9	17.5±0.9	12.9±0.8	33.6±0.8	29.4±0.8	30.8±0.6	28.8±0.7	12.5±0.3
CoMatch	12.7±0.7	9.5±0.5	14.8±0.8	10.3±0.4	28.5±0.6	26.4±0.6	28.8±0.7	25.8±0.5	8.8±0.9
MTC	20.4±0.9	13.5±0.8	21.8±1.2	14.3±0.6	36.7±0.9	30.9±0.5	36.9±1.2	29.6±0.6	13.6±0.7
OpenMatch	10.2±0.9	7.1±0.5	11.7±0.8	9.2±0.6	30.5±0.4	26.7±0.6	30.1±0.5	25.3±0.5	10.4±1.0
OSP	12.1±0.8	9.2±0.6	11.1±0.8	9.5±0.5	29.5±0.5	26.5±0.6	30.7±0.9	26.9±0.5	-
Ours	**9.6±0.7**	**5.8±0.4**	**8.1±0.9**	**6.8±0.5**	**26.6±0.3**	**23.6±0.2**	**26.4±0.6**	**23.8±0.4**	**6.1±0.6**

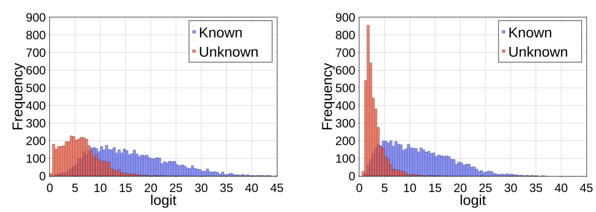
Table 2: Error rates with standard deviation for CIFAR-10/100 and ImageNet on 3 different folds.

Dataset	CIFAR-10				CIFAR-100				ImageNet-30
	Uncorr.		Corr.		Uncorr.		Corr.		
No. of labeled	50	100	50	100	50	100	50	100	10%
Labeled Only	65.8±0.6	67.7±0.5	63.9±0.6	64.7±0.6	71.3±0.8	74.1±0.8	64.4±1.0	68.7±0.9	81.3±1.0
FixMatch	54.2±0.6	58.5±0.4	55.9±0.4	59.4±0.5	64.1±0.8	66.7±0.5	60.4±0.7	62.6±0.5	87.9±0.6
CoMatch	47.6±0.5	47.7±0.6	48.1±0.6	48.9±0.6	60.1±1.3	61.7±1.2	55.6±0.5	57.9±0.9	65.8±1.2
MTC	96.5±0.4	98.2±0.3	73.5±0.6	78.2±0.5	81.7±2.8	82.6±3.4	69.3±2.5	70.4±3.5	93.8±0.8
OpenMatch	**97.9±0.4**	**99.6±0.3**	75.6±0.5	55.3±1.2	86.1±1.3	86.2±2.1	69.9±0.8	69.4±0.5	96.4±0.7
OSP	62.9±0.6	66.0±0.7	45.7±0.8	46.4±0.7	58.5±1.2	60.3±1.4	56.1±0.9	59.5±0.8	-
Ours	95.5±0.4	97.8±0.5	**90.2±0.7**	**96.4±0.3**	**87.9±0.8**	**91.8±0.6**	**78.7±0.7**	**81.4±0.5**	**97.4±0.5**

Table 3: AUROC performances of Table 2.

				Uncorr.		Corr.	
EMX	MHT	PLG	POG	Error	AUR.	Error	AUR.
				29.4	66.7	28.8	62.6
✓				29.3	82.2	30.2	73.3
✓	✓			27.5	89.1	28.2	77.7
✓		✓		25.8	63.9	25.8	60.1
✓		✓	✓	**23.5**	72.3	**23.6**	68.2
✓	✓	✓	✓	23.6	**91.8**	23.8	**81.4**

Table 4: Ablation studies on the individual modules.



(a) Single-head (b) Multi-head

Figure 5: Histogram plots of the logits estimated from models with single-head and multi-head structures. Red and blue bars correspond to the inliers and outliers, respectively.

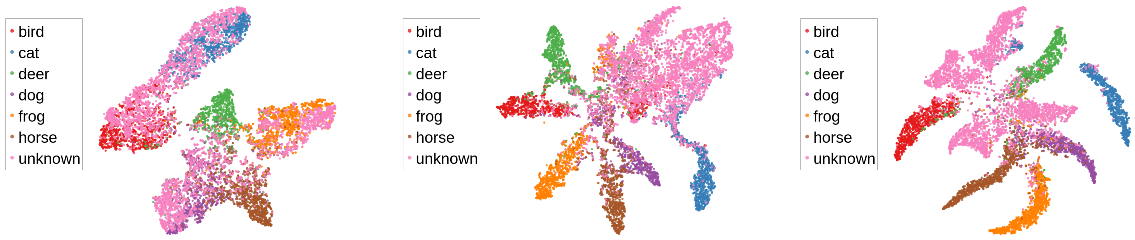
Following previous OSSL work (Saito, Kim, and Saenko 2021), we designate the first 20 classes alphabetically as known classes, and the rest as unknown. Only 10% of the known class data is labeled, with the remainder as unlabeled data. The results in the rightmost columns of Tables 2 and 3 demonstrate that our method, UAG, achieves state-of-the-art performance in terms of both error rates and AUROC.

Ablation Studies

To better understand the benefits of UAG, we conducted ablation studies on quantitative and qualitative evaluations. All experiments are performed on the test set of CIFAR datasets using models trained with 100 labeled data per each class.

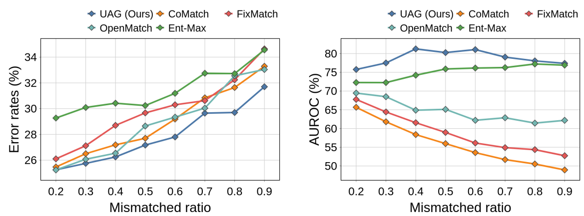
Effectiveness of the proposed components. Table 4 compares ablated models on the CIFAR-100 dataset. FixMatch serves as our baseline, with its results reported at the top. Entropy maximization alone (EMX) improves AUROC but not error rates. Multi-head training (MHT) significantly enhances both error rates and AUROC. Incorporating the pseudo-label graph (PLG) reduces error rates but worsens AUROC. Further addition of the pseudo-outlier graph (POG) achieves substantial progress in both metrics. UAG surpasses the baseline, showing notable improvements of 5.8% (5.0%) in error rates and 25.1% (18.8%) in AUROC over the baseline in the uncorrelated (correlated) setting.

Multi-Head Training Effectiveness. Fig. 5 shows the effect of exclusive multi-head training by comparing logit histograms between single- and multi-head structures. All the models are trained on uncorrelated settings. Multi-head training shows a clearer distinction in logit distribution between known and unknown classes compared to the single-head method, indicating the effectiveness of our approach in



(a) FixMatch (b) w/ Ent-Max (c) w/ UAG (Ours)

Figure 6: t-SNE visualization of embeddings obtained from the ablated models. Pink points denote outliers, while the other colored points represent distinct known classes. (a) A model trained solely with FixMatch. (b) A model trained with entropy maximization further applied on (a). (c) A model trained using the proposed method UAG.



(a) Error rates (b) AUROC

Figure 7: Results with various mismatched ratio of outliers.

Known / Unknown	40 / 60		80 / 20	
Method	Error	AUROC	Error	AUROC
Labeled Only	30.8	70.4	38.5	69.2
FixMatch	21.9	64.7	31.4	58.9
OpenMatch	19.7	69.1	28.9	66.5
Ours	**18.1**	**80.7**	**27.6**	**79.8**

Table 5: Results with the different number of known classes.

	Top-1 Acc.	Top-5 Acc.
Labeled Only	15.43	30.86
FixMatch	17.00	32.90
CoMatch	19.17	36.97
OpenMatch	18.65	35.68
Ent-Max	13.35	27.33
UAG (Ours)	**24.16**	**44.32**

Table 6: Top-1 and Top-5 accuracy for Semi-iNature 2021.

mitigating learning conflicts between outliers and inliers.

Effectiveness of the contrastive graph regularization. Fig. 6 demonstrates t-SNE visualization of feature distribution extracted from CIFAR-10 models trained in the correlated setting. Results reveal that a FixMatch-only model (a) struggles with inlier-outlier differentiation, whereas our proposed UAG (c) excels. Notably, our model identifies multiple centroids where outlier examples cluster, showcasing its robust discrimination compared to entropy maximization (b), which aims to collapse all outliers into a single cluster.

Further Analysis

Results on various mismatch ratio. We evaluate the robustness of our method against uncurated data corruption by analyzing performances across various mismatch ratios with the CIFAR-100 dataset (100 label per class) in the correlated setting. The results are depicted in Fig. 7. Across all mismatched scenarios, our UAG model consistently outperforms alternative approaches in terms of error rates and AUROC. These findings strongly affirm the effectiveness and superiority of our proposed method.

Results on the different number of known classes. Experiments were conducted with different numbers of known classes, and results are presented in Table 5. All the results are derived from CIFAR-100 (100 label per class) in the correlated setting. The table illustrates the consistently superior performance of our proposed method across all cases, regardless of the known and unknown class ratio.

Results on real-world dataset. We conducted additional experiments on the Semi-iNature-2021 dataset to verify the robustness of our framework. Employing ResNet-18 as our encoder, trained from scratch for 100 epochs, yielded results presented in Table 6. Our method demonstrates superior performance, surpassing existing baseline models even when applied to large-scale real-world datasets.

Conclusion

In this paper, we introduce unknown-aware graph regularization (UAG), a novel approach for open-set semi-supervised learning (OSSL). We concentrate on highly correlated scenarios, where inlier and outlier classes share the same superclass space, providing a challenging and realistic OSSL benchmark. Extensive experiments show UAG's outstanding performance across various OSSL scenarios. We consider our work a valuable baseline guiding future research in both semi-supervised learning (SSL) and OSSL.

Acknowledgements

This research was supported by the Challengeable Future Defense Technology Research and Development Program (912911601) of Agency for Defense Development in 2020 and was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant, funded by the Korea government (MSIT) (No. 2019-0-00079, Artificial Intelligence Graduate School Program (Korea University)).

References

Berthelot, D.; Carlini, N.; Cubuk, E. D.; Kurakin, A.; Sohn, K.; Zhang, H.; and Raffel, C. 2019a. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. In *Proc. International Conference on Learning Representations (ICLR)*.

Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; and Raffel, C. A. 2019b. Mixmatch: A holistic approach to semi-supervised learning. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*.

Cascante-Bonilla, P.; Tan, F.; Qi, Y.; and Ordonez, V. 2021. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 6912–6920.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual representations. In *Proc. International Conference on Machine Learning (ICML)*, 1597–1607.

Chen, Y.; Zhu, X.; Li, W.; and Gong, S. 2020b. Semi-supervised learning under class distribution mismatch. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 3569–3576.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Grandvalet, Y.; and Bengio, Y. 2004. Semi-supervised learning by entropy minimization. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*.

Guo, L.-Z.; Zhang, Z.-Y.; Jiang, Y.; Li, Y.-F.; and Zhou, Z.-H. 2020. Safe deep semi-supervised learning for unseen-class unlabeled data. In *Proc. International Conference on Machine Learning (ICML)*, 3897–3906.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

He, R.; Han, Z.; Lu, X.; and Yin, Y. 2022. Safe-Student for Safe Deep Semi-Supervised Learning with Unseen-Class Unlabeled Data. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 14585–14594.

Hendrycks, D.; and Gimpel, K. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proc. International Conference on Learning Representations (ICLR)*.

Hendrycks, D.; Mazeika, M.; Kadavath, S.; and Song, D. 2019. Using self-supervised learning can improve model robustness and uncertainty. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*.

Huang, J.; Fang, C.; Chen, W.; Chai, Z.; Wei, X.; Wei, P.; Lin, L.; and Li, G. 2021. Trash to treasure: harvesting OOD data with cross-modal matching for open-set semi-supervised learning. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 8310–8319.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*.

Kong, H.; Lee, G.-H.; Kim, S.; and Lee, S.-W. 2023. Pruning-Guided Curriculum Learning for Semi-Supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5914–5923.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. Department of Computer Science, University of Tronto.

Laine, S.; and Aila, T. 2016. Temporal ensembling for semi-supervised learning. In *Proc. International Conference on Learning Representations (ICLR)*.

Lee, D.-H.; et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop on challenges in representation learning*, 896.

Li, J.; Xiong, C.; and Hoi, S. C. 2021. Comatch: Semi-supervised learning with contrastive graph regularization. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 9475–9484.

Liang, S.; Li, Y.; and Srikant, R. 2017. Enhancing the reliability of out-of-distribution image detection in neural networks. In *Proc. International Conference on Learning Representations (ICLR)*.

Liu, W.; Wang, X.; Owens, J.; and Li, Y. 2020. Energy-based out-of-distribution detection. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*.

Maeng, H.; Liao, S.; Kang, D.; Lee, S.-W.; and Jain, A. K. 2013. Nighttime face recognition at long distance: Cross-distance and cross-spectral matching. In *Proc. Asian Conference on Computer Vision (ACCV)*, 708–721.

Miyato, T.; Maeda, S.-i.; Koyama, M.; and Ishii, S. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8): 1979–1993.

Nam, W.-J.; Gur, S.; Choi, J.; Wolf, L.; and Lee, S.-W. 2020. Relative attributing propagation: Interpreting the comparative contributions of individual units in deep neural networks. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2501–2508.

Oliver, A.; Odena, A.; Raffel, C. A.; Cubuk, E. D.; and Goodfellow, I. 2018. Realistic evaluation of deep semi-supervised learning algorithms. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*.

Saito, K.; Kim, D.; and Saenko, K. 2021. Openmatch: Open-set consistency regularization for semi-supervised learning with outliers. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*.

Sajjadi, M.; Javanmardi, M.; and Tasdizen, T. 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*.

Su, J.-C.; and Maji, S. 2021. The semi-supervised inaturalist challenge at the fgvc8 workshop. In *arXiv preprint arXiv:2106.01364*.

Sun, Y.-X.; and Wang, W. 2022. Exploiting mixed unlabeled data for detecting samples of seen and unseen out-of-distribution classes. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 8386–8394.

Tack, J.; Mo, S.; Jeong, J.; and Shin, J. 2020. Csi: Novelty detection via contrastive learning on distributionally shifted instances. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*.

Wang, H.; Li, Z.; Feng, L.; and Zhang, W. 2022. Vim: Out-of-distribution with virtual-logit matching. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4921–4930.

Wang, Y.; Qiao, P.; Liu, C.; Song, G.; Zheng, X.; and Chen, J. 2023. Out-of-Distributed Semantic Pruning for Robust Semi-Supervised Learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 23849–23858.

Xie, Q.; Dai, Z.; Hovy, E.; Luong, T.; and Le, Q. 2020. Unsupervised data augmentation for consistency training. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*.

Yu, Q.; Ikami, D.; Irie, G.; and Aizawa, K. 2020. Multi-task curriculum framework for open-set semi-supervised learning. In *Proc. European Conference on Computer Vision (ECCV)*, 438–454.

Zagoruyko, S.; and Komodakis, N. 2016. Wide Residual Networks. In *Proc. British Machine Vision Conference (BMVC)*.

Zheng, M.; You, S.; Huang, L.; Wang, F.; Qian, C.; and Xu, C. 2022. Simmatch: Semi-supervised learning with similarity matching. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 14471–14481.