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

2024 AAAI Conference on Artificial Intelligence

Heejo Kong, Suneung Kim, Ho-Joong Kim and Seong-Whan Lee
Korea University
{hj_kong, se_kim, hojoong_kim, sw.lee}@korea.ac.kr

Background

- **Semi-supervised learning (SSL)**
  - ➢ One of the most famous approach for data-efficient learning
  - ➢ Leveraging abundant unlabeled data for learning with only limited labeled data
- **An optimistic assumption of existing SSL algorithms**
  - ➢ Assumption: labeled and unlabeled data are drawn from the identical class distribution
    - ✓ In other words, all training data consists of only known k-way classes
  - ➢ However, in practical scenarios, this optimistic assumption can be easily violated
    - ✓ Unlabeled dataset often includes out-of-class data
    - ✓ **<u>This uncurated unlabeled data severely degrade the performance of existing SSL methods</u>**

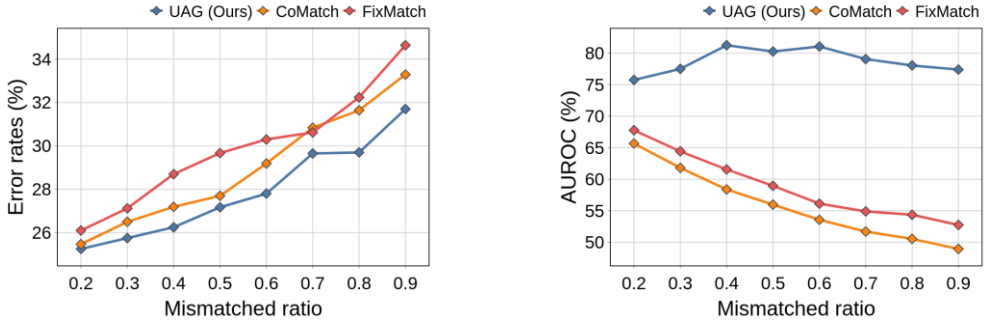

Figure 1. Error rates(left) and AUROC(right) results with various mismatched ratio of outliers. For detailed explanation, please refer to Figure 7 in the main paper.

Background
Motivations
Our Approach
Experiments

Motivations

- **Open-set semi-supervised learning (OSSL)**
  - ➢ Unlike conventional SSL, OSSL considers the existence of out-of-class data in unlabeled set
    - ✓ OSSL aims to train a model that not only classifies samples from known categories, i.e., inliers, but also identifies samples from novel classes as outliers
  - ➢ While OSSL is more realistic and practical, it has been rarely considered in previous literatures
- **Similarity-based OSSL approach**
  - ➢ Recent notable works 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
  - ➢ However, similarity-based measures can **easily fail to detect highly correlated outliers**

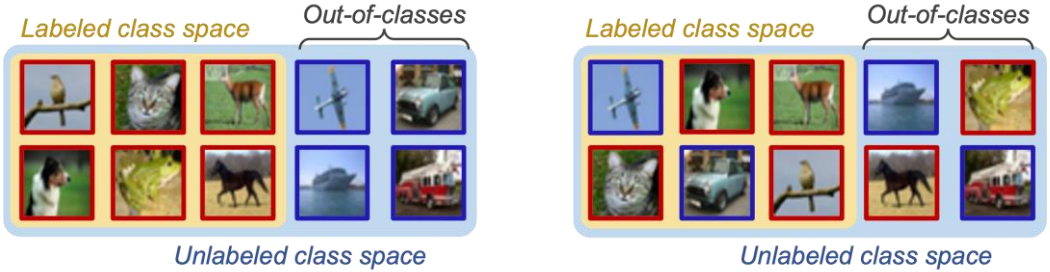

Figure 2. An illustration of the example for uncorrelated(left) and correlated outliers(right).
The classes marked with red and blue box share the same superclass, animal and transportation, respectively

Motivations

- **Uncertainty-based OSSL approach**
  - ➢ As an alternative to the feature similarity, some OSSL works exploit uncertainty in logit space
    - ✓ In contrast to the features exclusively disregard the classification weights with class-dependent information, the logits effectively activate task-specific information related to high-level semantic attributes
    - ✓ This property enables uncertainty-based measures to serve as a robust detector for highly correlated outliers
  - ➢ However, the performance of uncertainty-based measures is still not competitive in practice

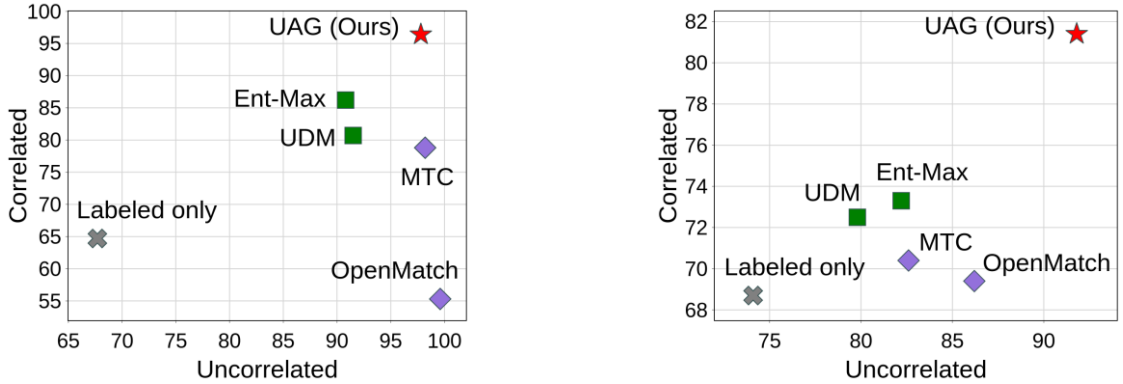

Figure 3. AUROC of five OSSL algorithms trained on CIFAR-10(left) and CIFAR-100(right) datasets with correlated and uncorrelated settings. Methods marked with square use the uncertainty in logits; methods with diamond use the similarity in features

Our Approach

- **(Baseline) Uncertainty-based outlier detection**
  - ➢ We exploit the uncertainty of logits to construct a baseline outlier detector
    - ✓ Adopting a maximum softmax probability (MSP) as an uncertainty score
    - ✓ Considering the samples with the low scores as outliers
      - The samples 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 two-component GMM with EMA updates

- MSP scores

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

- EMA thresholds

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

- **Reasons for under-performing of uncertainty-based OSSL approach**
  - ➢ We assert that the under-performing of uncertainty-based methods mainly originated from two aspects:
    - ✓ **Aspect 1)** 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
    - ✓ **Aspect 2)** Although the outliers are composed of multiple novel classes, previous works train the model by assigning them into one generic class, unknown

Our Approach

- **Unknown-Aware Graph Regularization (UAG)**
  - ➢ UAG aims to enhance the uncertainty-based OSSL framework by addressing the introduced two aspects
  - ➢ The key components of the proposed UAG consist of the followings:
    - ✓ Exclusive multi-head training for addressing an aspect 1
    - ✓ Contrastive graph regularization for addressing an aspect 2
      - Overall loss function

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

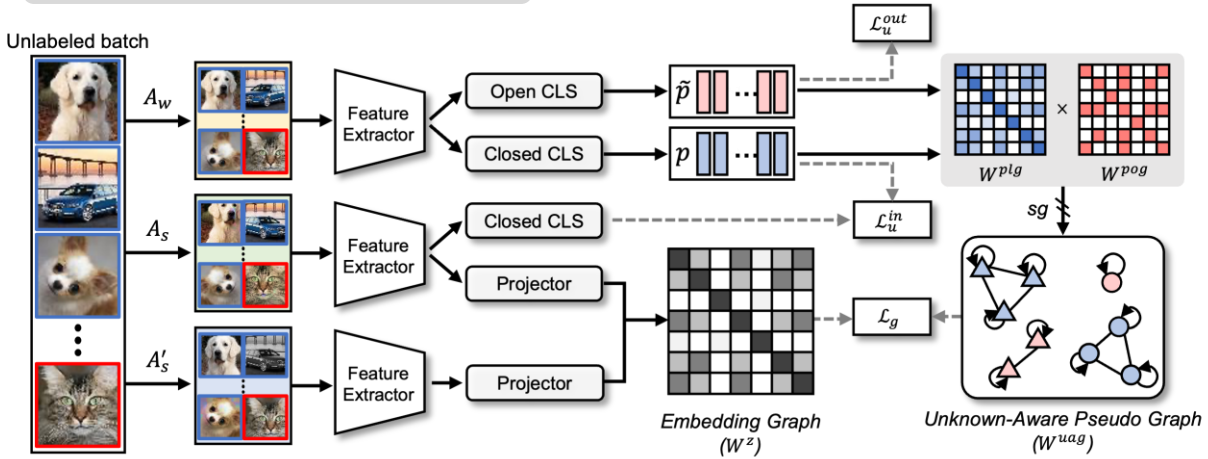

Figure 4. Overall framework of the proposed UAG. For detailed explanation, please refer to Figure 3 in the main paper.

Our Approach

- **Exclusive multi-head training**
  - ➢ Statistical rationale of the conflict between training objectives for inliers and outliers
    - ✓ Note that existing uncertainty-based methods only depend on a single classifier head in their training
      - A single classifier head trains to maximize entropy of pseudo-outliers and minimize entropy of pseudo-inliers
    - ✓ This training strategy allows the model to learn a significant amount of erroneous predictions
      - These false predictions act as noise that conflicts with our intended target objective, as they are learned to minimize entropy even when outlier entropy should be maximized, and vice versa

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

Figure 5. Results of the models trained solely on labeled data for CIFAR-10 (100 labeled per class)

  - ➢ Whereas, in our method, two classifier heads learn the objectives for inliers and outliers independently
    - ✓ While both classifiers are learned on labeled data, pseudo-inliers and -outliers are learned independently
    - ✓ **<u>It can inherently prevent the conflict that arises when a single classifier learns both objectives</u>**

- Supervised loss for labeled data

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

- Unsupervised loss for unlabeled data

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

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

Our Approach

- **Contrastive graph regularization**
  - ➢ Underlying rationale of the convergence of all outliers into a single representative space
    - ✓ In theoretical terms, the weights function as representative points for the embedding features z of each class
      - When considering updates only for the weights $\omega \in \mathbb{R}^{k \times d}$ of the last classifier in learning process
      - By minimizing the cross-entropy loss for inliers, $H(y, f(x,\theta);\omega) = -y \log \sigma(f(x,\theta) \cdot \omega^T)$, the weights converge to points in the embedding space that maximize the similarity with samples corresponding to the k-th label
    - ✓ 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 outliers converge to a single cluster orthogonal to weights ω
  - ➢ An interesting observation of conventional SSL without entropy maximization
    - ✓ An existing SSL method, FixMatch, has 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**

- Objectives for entropy minimization

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

- Objectives for entropy maximization

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

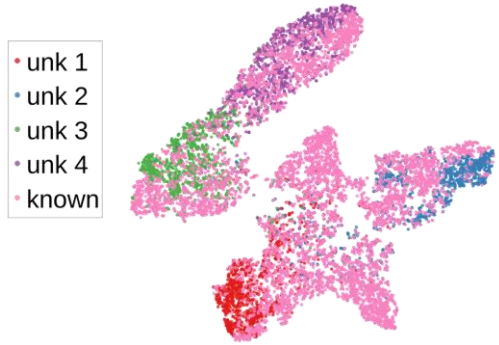

Figure 6. t-SNE visualization of embeddings obtained from a model trained with FixMatch

Our Approach

- **Contrastive graph regularization**
  - ➢ Based on the above empirical rationale, we derive an unknown-aware graph regularization that allows the outliers to form multiple clusters in their embeddings
    - ✓ Building a pseudo-label graph by leveraging the batch-wise predictions of a closed-set classifier
      - Each sample is connected to itself with the strongest edge of value 1 as self-loop, while the samples with similarity less than τ_g are not connected.
    - ✓ Building a pseudo-outlier graph by leveraging the batch-wise predictions of an open-set classifier
      - Note that $\eta_b = (s_b < \hat{\tau}^{out})$ is an outlier indicator, and returns 1 for values predicted as outliers and 0 for the rest
    - ✓ An unknown-aware pseudo graph is obtained by matrix multiplication of two graphs; $W^{uag} = W^{plg} \cdot W^{pog}$
      - Pseudo-label graph

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

      - Pseudo-outlier graph

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

    - ✓ Constructing the embedding graph based on the batch-wise similarity of the projected embedding
    - ✓ Then, the graph contrastive loss is derived to minimize the cross-entropy between the two normalized graphs
      - Embedding graph

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

      - Graph contrastive loss

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

Experiments

- **Experimental results**
  - ➢ Datasets: CIFAR-10 & CIFAR-100, ImageNet-30

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 1. Error rates with standard deviation for CIFAR-10/100 and ImageNet-30 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 2. AUROC performance of Table 1

Experimental Results

- **Ablation studies**
  - ➢ Effectiveness of the proposed components (results on Table 3)

				Uncorr.		Corr.	
Ent-Max	MHT	PLG	POG	Error	AUROC	Error	AUROC
				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 3. Ablation studies on the individual modules.

  - ➢ Effectiveness of the multi-head training (results on Figure 7)

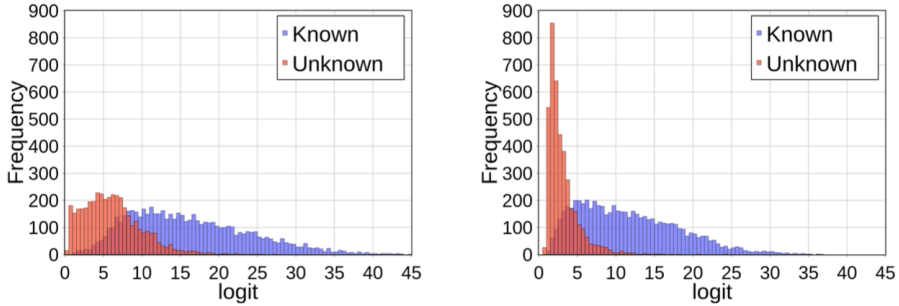

Figure 7. Histogram plots of the logits estimated from models with single-head and multi-head structures. Red and blue bars correspond to the inliers and outliers, respectively.

Experimental Results

- **Ablation studies**
  - ➢ Effectiveness of the contrastive graph regularization (results on Figure 8)

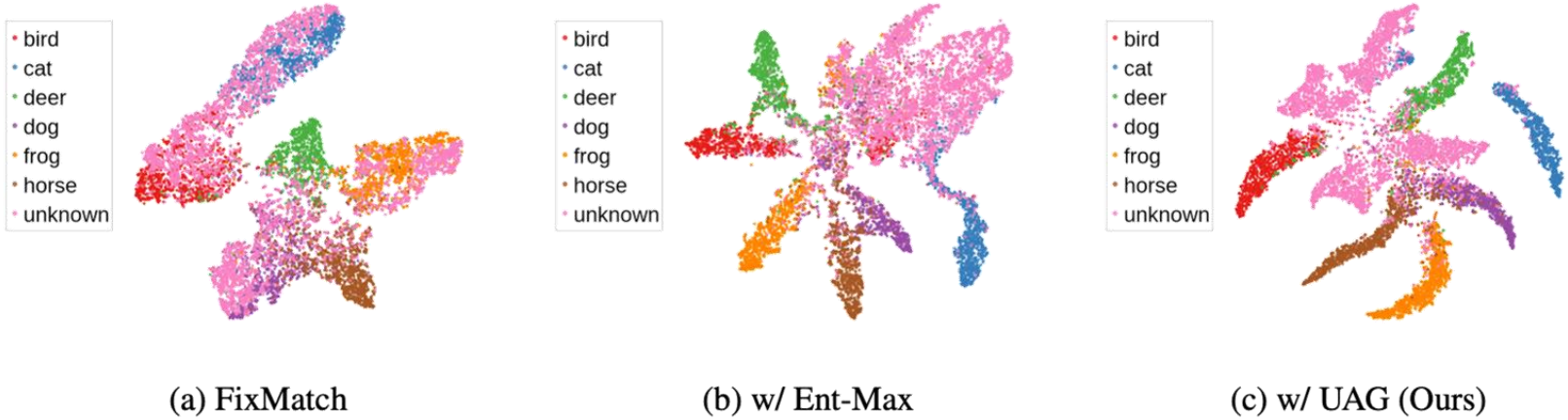

Figure 8. t-SNE visualization of embeddings obtained from the ablated models. Pink points denote outliers, while the other colored points represent distinct known classes. (a) A model trained solely with FixMatch. (b) A model trained with entropy maximization further applied on (a). (c) A model trained using the proposed method UAG.

Experimental Results

- **Further analysis**
  - Results on various mismatch ratio (results on Figure 9)

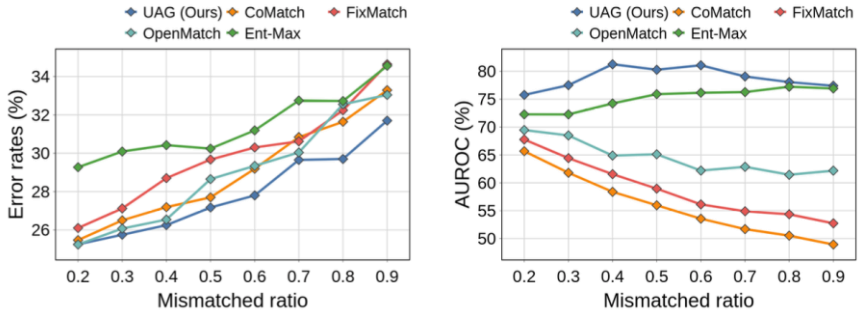

Figure 9. Error rates(left) and AUROC(right) results with various mismatched ratio of outliers.

  - Results on different number of known classes (results on Table 4)
  - Results on real-world dataset (results on Table 5)

Method	40 known / 60 unknown		80 known / 20 unknown	
	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 4. 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 5. Top-1 and Top-5 accuracy for Semi-iNature 2021.

Thank you for your attention!

Heejo Kong, Suneung Kim, Ho-Joong Kim and Seong-Whan Lee

Korea University

{hj_kong, se_kim, hojoong_kim, sw.lee}@korea.ac.kr