Superclass Learning with Representation Enhancement

CVPR 2023

THU-PM-330

Zeyu Gan, Suyun Zhao, Jinlong Kang, Liyuan Shang, Hong Chen, Cuiping Li

THU-PM-330

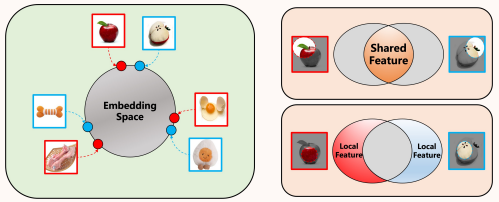
Superclass Learning with Representation Enhancement

Zeyu Gan, Suyun Zhao, Jinlong Kang, Liyuan Shang, Hong Chen, Cuiping Li

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

1 Introduction & Motivation

➢ **Definition:** In real-world applications, the criteria for image classification are often determined by human cognition rather than the images' features. In some scenarios, a category of images may contain various subclasses due to overly coarse-grained criteria, resulting in a lack of common semantic features even among images belonging to the same class. For example, in the context of waste classification, images of recyclable waste include a wide range of items, from beverage cans to books, with no apparent commonality. This article defines this phenomenon as the ***Superclass Learning*** problem.



(a) Image Feature Domain (b) Separate Fine Domain (c) Superclass Concept Domain

Illustration of Superclass Learning

➢ **Characteristics of Superclass Learning**

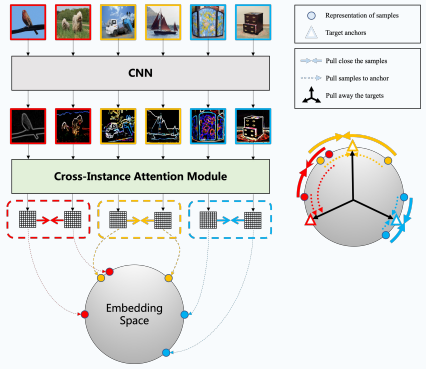
- Subclasses within a superclass are usually scattered and share few common features.
- Instances from different superclasses may have common features, leading to potential confusion between classes.

➢ **Challenges of Superclass Learning**

- **Breaking the original basic class decision boundaries**(Fig 1.b): It is necessary to divide the domain into smaller, more meaningful domains that better represent the actual problem (e.g., splitting the apple domain into fruit and toy domains).
- **Reconstructing decision boundaries at the superclass level**(Fig 1.c): The goal is to merge relevant class domains into a new superclass domain that accurately represents the intended classification (e.g., combining fruit apples, eggs, and bones into a food waste superclass domain).

2 Method Framework

We propose a novel representation enhancement method (SCLRE) to address superclass learning problem.



Overview of the Framework of SCLRE

SCLRE processes the images in the following steps.

1) **Encoding.** The images generate their representations through CNN.
2) **Enhancement.** Then they mix with each other in a trainable cross-instance attention module for enhancement.
3) **Adjustment.** Enhanced representations are then adjusted according to their superclass labels and the target anchors.

➢ **Cross-Instance Attention Module**

We calculate the attention weights by putting the batch images into a cross-instance attention module. Then we use the attention weights to mix the representations with each other.

➢ **SCLRE Adjustment Loss**

We adjust the enhanced representations under the guidance of a supervised contrastive loss. And we also preset target anchors for each superclass to help the adjustment process.

3 Experiment Result

We integrate the initial class labels of the original dataset to superclass labels, according to field knowledge. We conduct extensive experiments on several artificially constructed superclass datasets to demonstrate the effectiveness of our proposed approach SCLRE.

Dataset	Method	Accuracy(%)
CIFAR100-3	Baseline	72.8
	SupCon [22]	78.1
	SimCLR [4]	79.0
	SCLRE	**80.1**
CIFAR100-4	Baseline	76.0
	SupCon [22]	80.1
	SimCLR [4]	80.6
	SCLRE	**84.0**
CIFAR100-7	Baseline	68.9
	SupCon [22]	72.7
	SimCLR [4]	73.9
	SCLRE	**78.1**

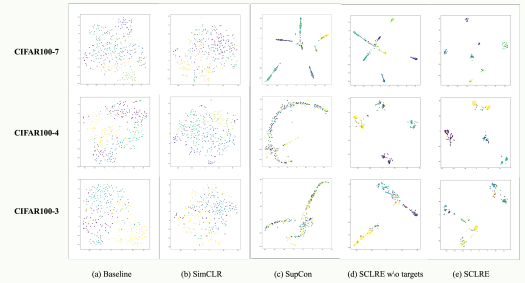
Table 1. **Classification accuracy on low pixels dataset: CIFAR100.** We compared the classification accuracy on CIFAR100-3, CIFAR100-4, and CIFAR100-7 datasets.

	mini-Imagenet [38]	FMoW [23]	Adience [9]	VOC [10]
Baseline	87.7	56.5	65.7	78.3
SupCon	87.0	58.3	66.2	78.2
SimCLR	**90.1**	51.5	52.9	80.0
SCLRE	89.3	**64.8**	**68.7**	**81.5**

Table 2. **Classification accuracy on high-resolution datasets.** We compared the accuracy on datasets with more complex and informative content.

➢ **Classification Accuracy**

We calculate the accuracy of SCLRE and other compared methods. The results show that SCLRE outperforms other SOTA representation learning frameworks.



Visualization of Representation Space

➢ **Representation Visualization**

We use t-SNE to visualize the representation space of each compared method. The result shows that SCLRE can effectively generate superclass-aware representations and make the boundaries of each superclass more clear.

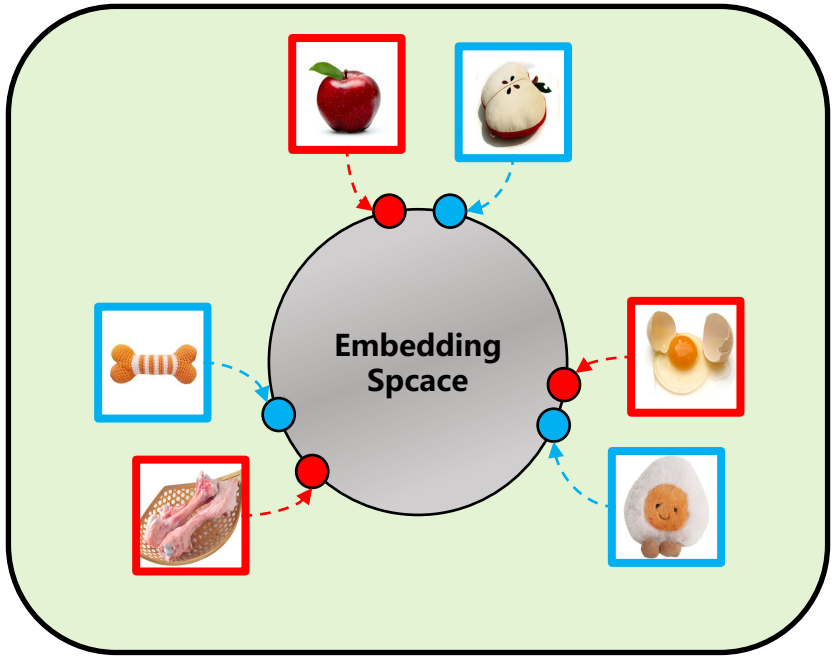
4 Generalization Analysis

We prove that the generalization error of SCLRE can be bounded by the similarity of attention vectors from the same superclass in the form of the following equation. During the training process, the sample pairs in O_2 can naturally have high similarity attention vectors, thereby reducing the upper bound of the generalization error, which is consistent with the phenomenon we observed in the experiment.

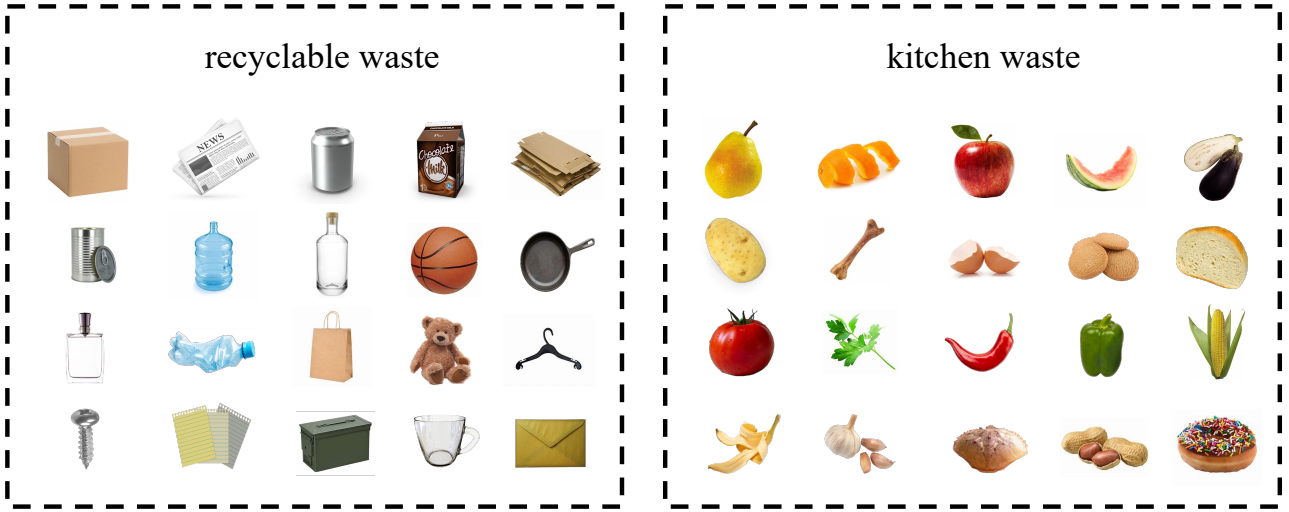
➢ **Upper Bound of Generalization Error:**

$$Err(G_f) \leq (1-\sigma) + \sqrt{2}\eta(\epsilon)\sqrt{1 - C_\varphi - \frac{(1-\rho)K}{M_1 M_2}\mathbb{E}_{v_1,v_2 \in O_2} s(a_1, a_2)}$$

Introduction & Motivation



Lack of Common Features

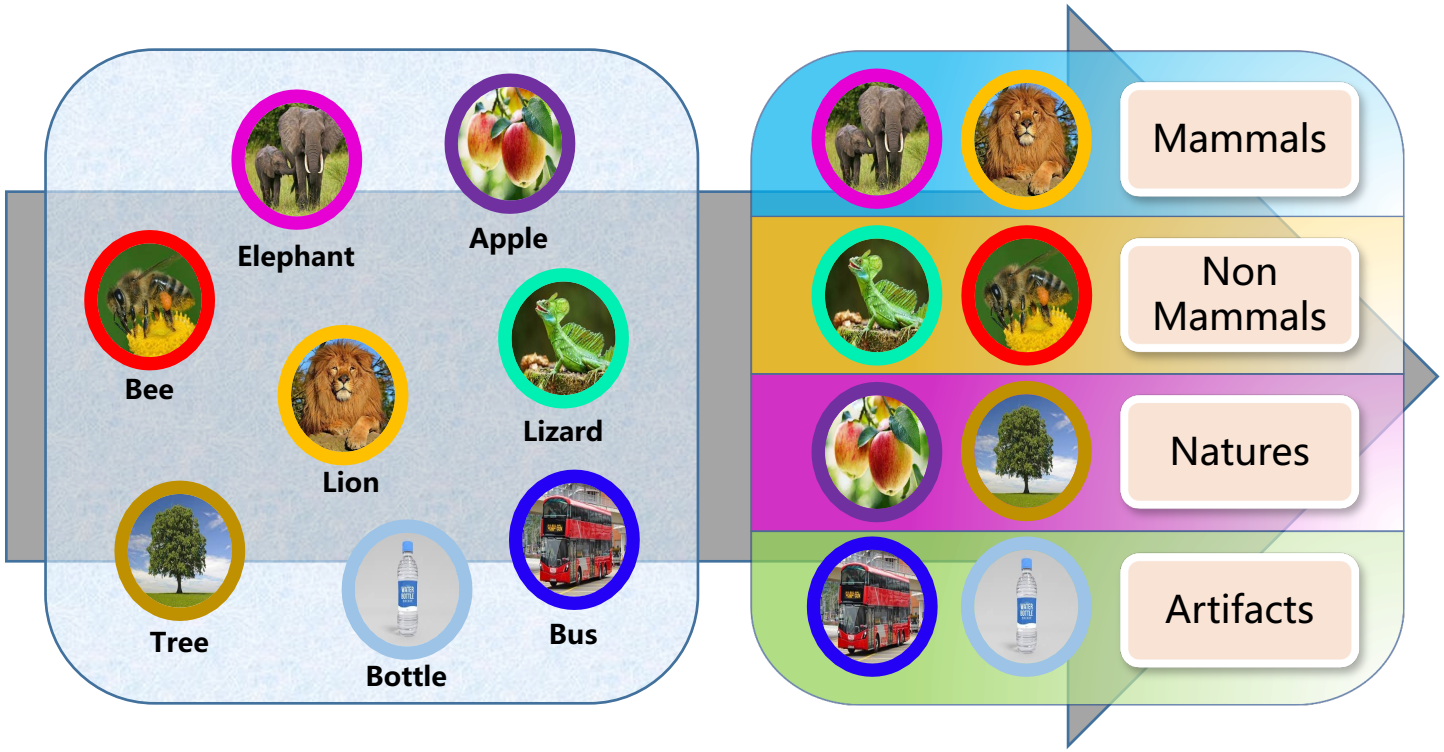


To Learn Representations: Instances with similar image features will be closer in the representation space

Real-World Classification: Classification criteria do not depend on image features, but are artificially defined

Image Feature ⇔ Classification Criteria

Introduction & Motivation

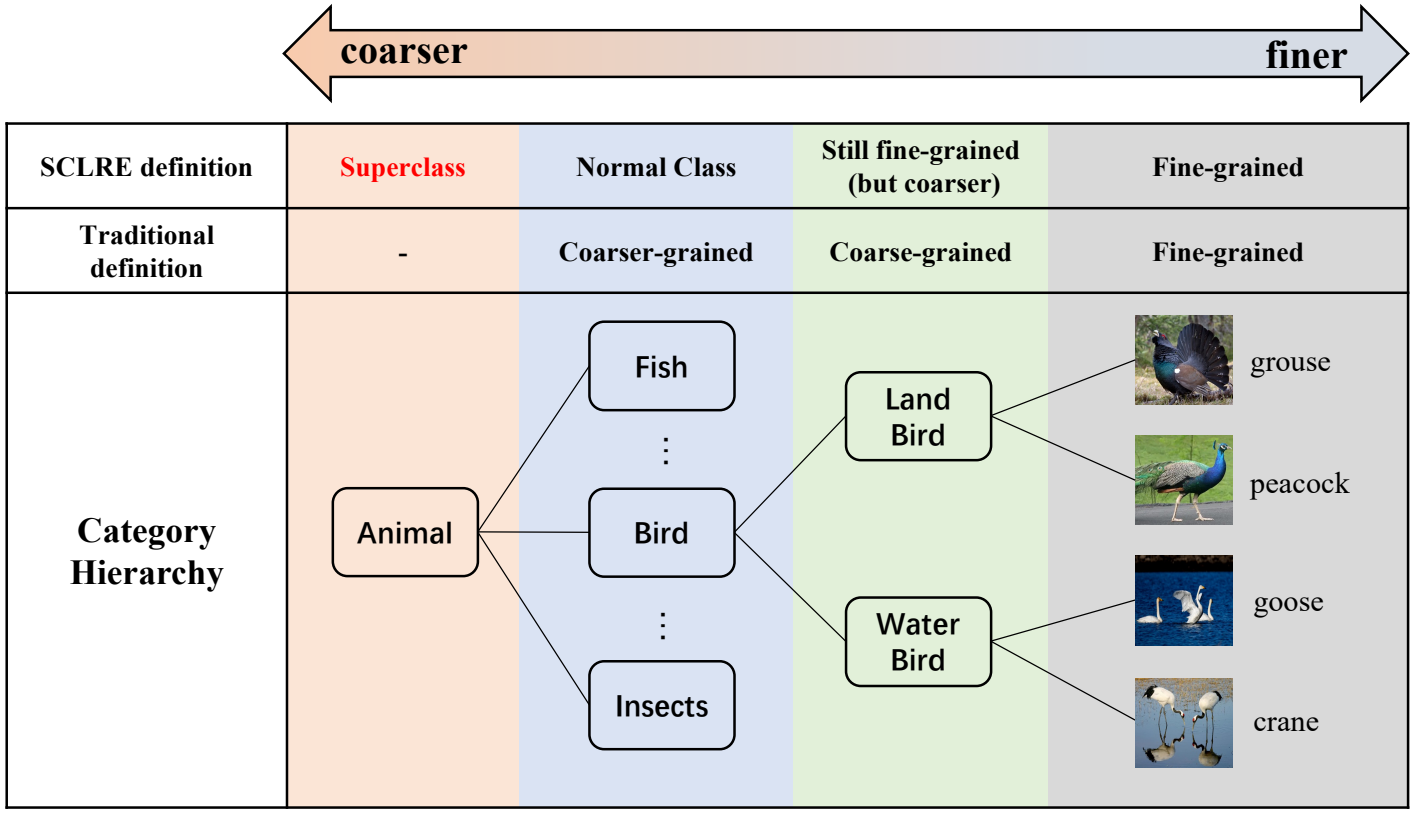


Characteristics of Superclass Learning

- **Scattered raw classes:** Subclasses within a superclass are usually scattered and share few common features.
- **Unclear Boundary:** Instances from different superclasses may have common features, leading to potential confusion between classes.

Definition: We call a class with such characteristics a *Superclass*.

Related Work



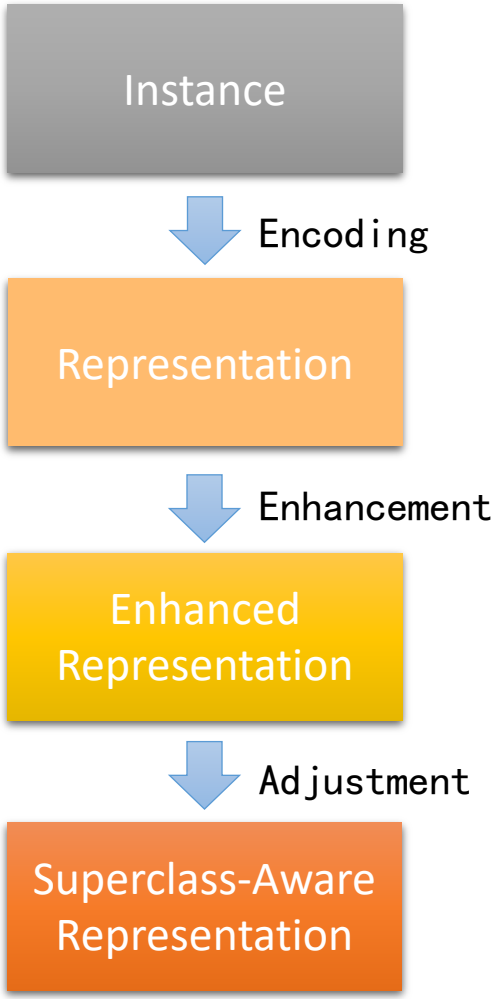
Existing studies focused more on **fine-grained** field, while less attention was paid to **coarse-grained** field.

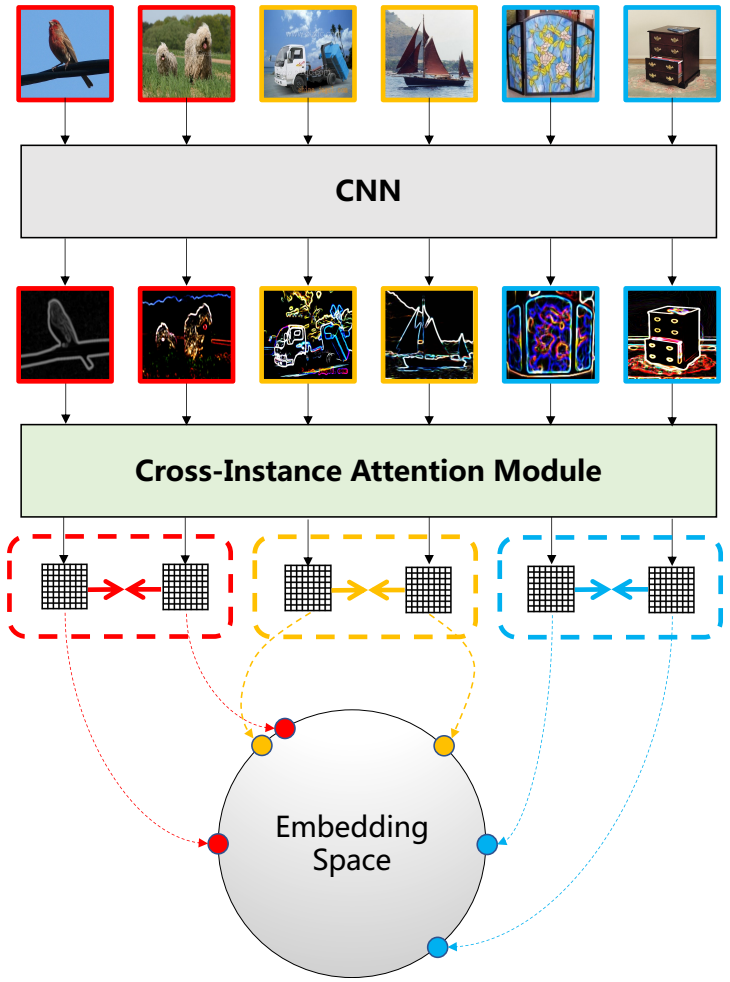
Challenges of Superclass Learning:
Difficult to extract representations at the superclass level, resulting in poor model performance.

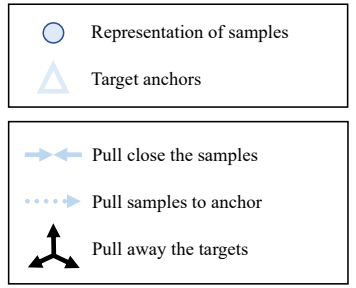
Need a new method to extract superclass-aware representations

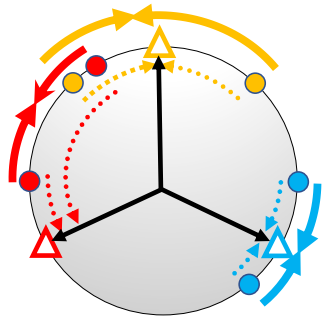
Method

We propose a novel representation enhancement method (SCLRE) to address superclass learning.

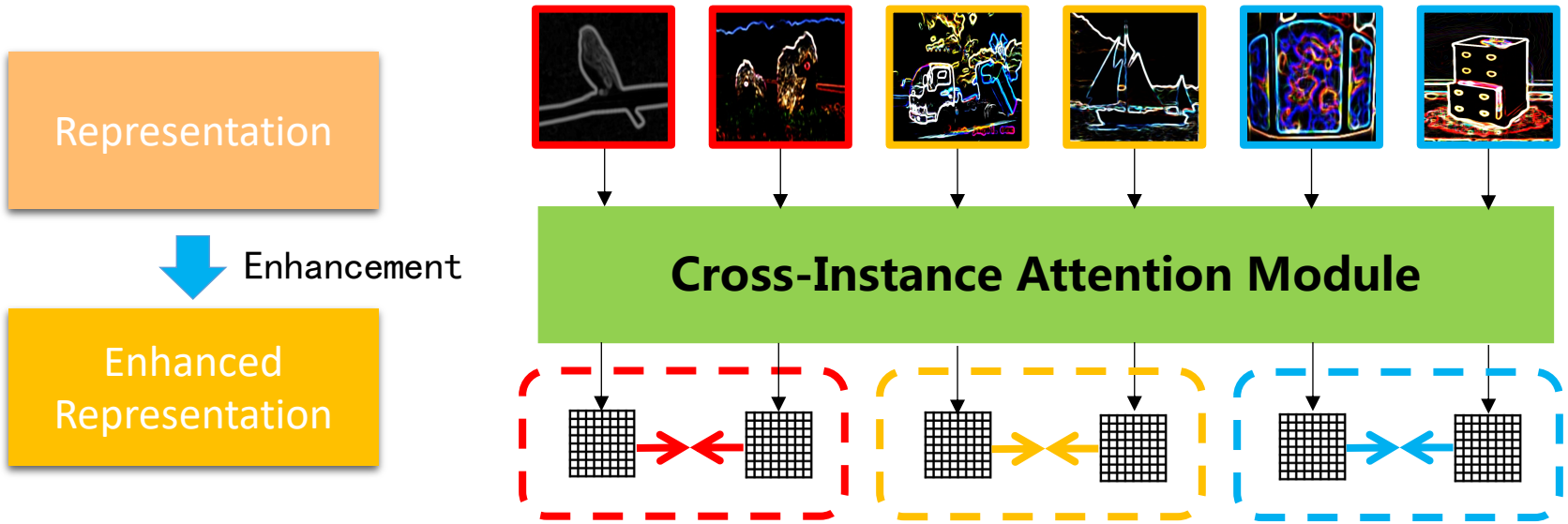








Method–Enhancement



Cross-Instance Attention Module

We use Cross-Instance Attention (CIA) Module to enhance representations.

1. **Calculate Attention:** We calculate the attention weights by putting the batch images into a cross-instance attention module.
2. **Enhancement:** We use the attention weights to mix the representations with each other.

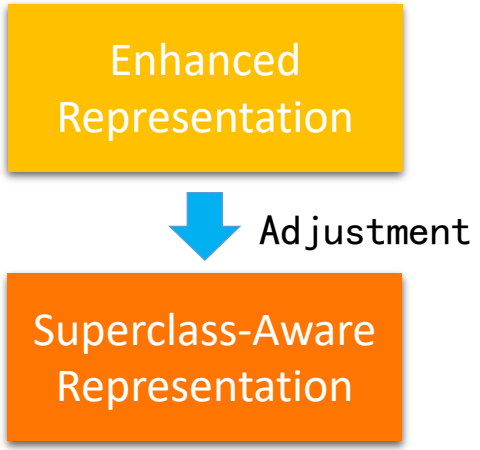
Attention Matrix

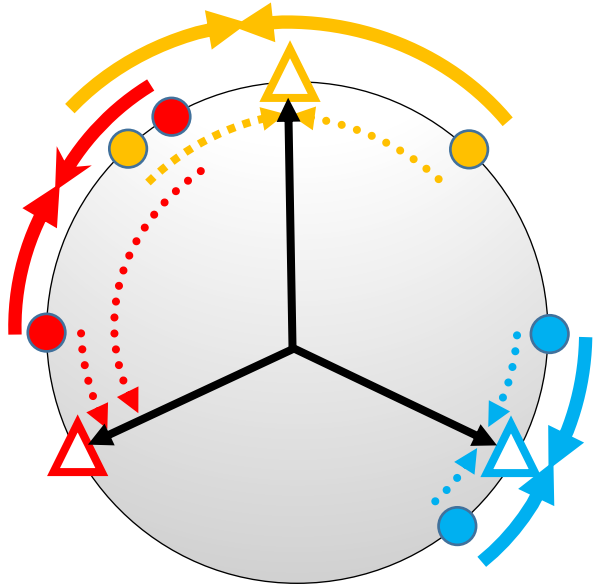
$$\mathrm{ADM}(Z) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right).$$

Enhancement

$$v_j = \mathrm{EnH}(z_j; Z) = a_j \times Z,$$

Method-Adjustment





Adjustment Loss

After enhancement, the representations are then adjusted according to their superclass labels and the target anchors.

1. **Contrastive Adjustment Loss:** We adjust the enhanced representations under the guidance of a supervised contrastive loss.
2. **Targeted Adjustment Loss:** We also preset target anchors for each superclass to help the adjustment process.

Contrastive Adjustment Loss

$$\ell(v_i, v^+) = -\log \frac{\exp(s(v_i, v^+)/\tau)}{\exp(s(v_i, v^+)/\tau) + \sum_{v^- \notin P(v_i)} \exp(s(v_i, v^-)/\tau)}.$$

$$\mathcal{L}_{ca} = \sum_{i=1}^{N} \frac{1}{||P(v_i)|| - 1} \sum_{v^+ \in P(v_i) \setminus \{v_i\}} \ell(v_i, v^+),$$

Targeted Adjustment Loss

$$\mathcal{L}_{ta} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(s(v_i, t(v_i))/\tau)}{\sum_{t \in T} \exp(s(v_i, t)/\tau)},$$

Method-Adjustment

SCLRE Loss

We also keep the cross-entropy loss to prevent model from collapsing. Then, we use hyperparameters α and λ to combine the losses.

$$\mathcal{L}_{SCLRE} = (1-\alpha)\mathcal{L}_{ce} + \alpha\mathcal{L}_{ca} + \lambda\mathcal{L}_{ta},$$

Cross-Entropy

$$\mathcal{L}_{ce} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{i,k}\log M(v_i)_k,$$

Contrastive Adjustment Loss

$$\ell(v_i, v^+) = -\log\frac{\exp(s(v_i,v^+)/\tau)}{\exp(s(v_i,v^+)/\tau)+\sum\limits_{v^-\notin P(v_i)}\exp(s(v_i,v^-)/\tau)}.$$

$$\mathcal{L}_{ca} = \sum_{i=1}^{N}\frac{1}{||P(v_i)||-1}\sum_{v^+\in P(v_i)\setminus\{v_i\}}\ell(v_i,v^+),$$

Targeted Adjustment Loss

$$\mathcal{L}_{ta} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(s(v_i,t(v_i))/\tau)}{\sum_{t\in T}\exp(s(v_i,t)/\tau)},$$

Experiment

We reorganize the raw classes of these datasets and make them superclass problems. The classification results show that SCLRE outperforms the SOTA representation learning framework.

Dataset	Method	Accuracy(%)
CIFAR100-3	Baseline	72.8
	SupCon [22]	78.1
	SimCLR [4]	79.0
	SCLRE	**80.1**
CIFAR100-4	Baseline	76.0
	SupCon [22]	80.1
	SimCLR [4]	80.6
	SCLRE	**84.0**
CIFAR100-7	Baseline	68.9
	SupCon [22]	72.7
	SimCLR [4]	73.9
	SCLRE	**78.1**

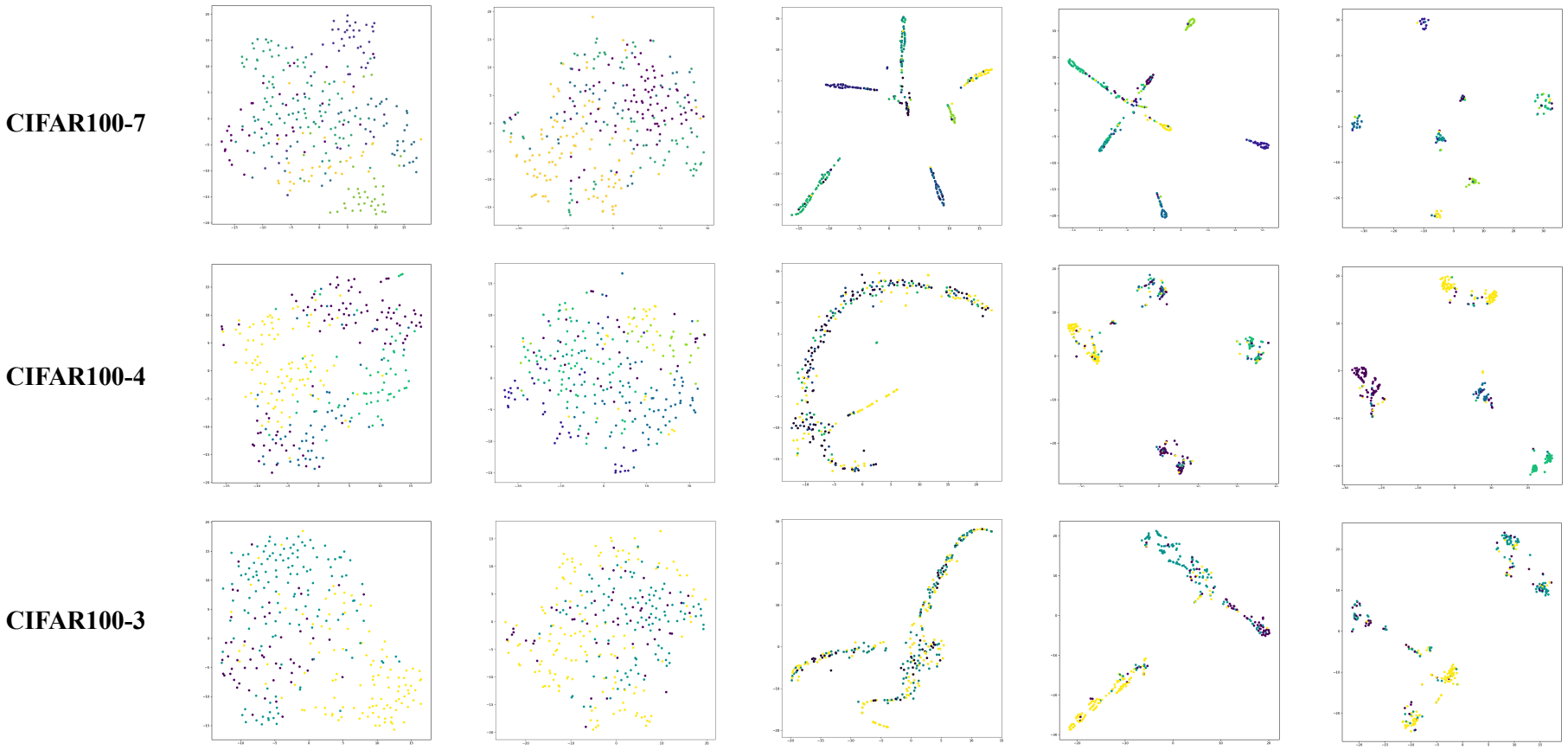
Table 1. **Classification accuracy on low pixels dataset: CIFAR100.** We compared the classification accuracy on CIFAR100-3, CIFAR100-4, and CIFAR100-7 datasets.

	mini-Imagenet [38]	FMoW [23]	Adience [9]	VOC [10]
Baseline	87.7	56.5	65.7	78.3
SupCon	87.0	58.3	66.2	78.2
SimCLR	**90.1**	51.5	52.9	80.0
SCLRE	89.3	**64.8**	**68.7**	**81.5**

Table 2. **Classification accuracy on high-resolution datasets.** We compared the accuracy on datasets with more complex and informative content.

Experiment

We use t-SNE to visualize the representation space of each compared method. The result shows that SCLRE can effectively extract superclass-aware representations and make the boundaries of each superclass more clear.



(a) Baseline (b) SimCLR (c) SupCon (d) SCLRE w\o targets (e) SCLRE

Experiment

We test the sensitivity of hyperparameters α, λ and Batch Size on the CIFAR100-7 dataset.

α	0	0.1	0.3	0.5	0.7	1
Accuracy	69.2	72.1	73.7	**75.1**	73.6	-

λ	0	0.1	0.2	0.4	0.8	1
Accuracy	75.1	76.7	77.9	76.9	**78.1**	76.0

Batch Size	64	128	256	512
Accuracy	76.7	76.9	77.3	**77.9**

Conclusion:

1. α and λ have peaks that make the model perform better.

2. The increase of *Batch Size* has improved the performance of the model, which may be due to the fact that a larger Batch brings more samples, so there are more enhancement options.

Experiment

We performed an ablation analysis between segments on SCLRE and explored their impact on model performance on CIFAR100-7.

Architecture	Params.(M)	FLOPs(G)	Acc.(%)
Baseline	23.51	1.30	68.9
+CIA	37.21	1.31	69.2
+$\mathcal{L}_{ca}$	-	-	72.7
+CIA, $\mathcal{L}_{ca}$	-	-	75.1
SCLRE (+CIA,$\mathcal{L}_{ca}$,$\mathcal{L}_{ta}$)	37.21	1.32	**77.9**

We replaced different Backbone and explored the robustness of SCLRE on CIFAR100-3.

	ResNet-18	ResNet-34	ResNet-50	ResNet-101	ResNet-152
Baseline	83.9	83.7	78.2	83.4	82.6
SCLRE	86.8	87.6	82.9	87.7	87.2
Improve.(%)	**+2.9**	**+2.9**	**+4.7**	**+4.3**	**+4.6**

Analysis of Generalization

We prove that the generalization error of SCLRE can be bounded by the similarity of attention vectors from the same superclass in the form of the following equation.

Contrastive Adjustment Loss

$$\mathcal{L}_{ca} \propto -\mathbb{E}_{v,v'}\mathbb{E}_{\substack{v_1,v_2\in P(v)\\ v^-\in P(v')}} \log \frac{\exp(v_1^T v_2)}{\exp(v_1^T v_2)+\exp(v_1^T v^-)}$$

$$= \underbrace{-\mathbb{E}_v \mathbb{E}_{v_1,v_2\in P(v)} v_1^T v_2}_{L_1}$$

$$+ \underbrace{\mathbb{E}_{v,v'}\mathbb{E}_{\substack{v_1,v_2\in P(v)\\ v^-\in P(v')}} \log\left(\exp(v_1^T v_2)+\exp(v_1^T v^-)\right)}_{L_2}$$

By splitting *Contrastive Adjustment Loss*, it can be rewritten in two parts: $L_1 + L_2$

Analysis of Generalization

Lemma 1

The upper bound of the generalization error of classifier G_f is determined by L_1

$$Err(G_f) \leqslant (1-\sigma) + \eta(\epsilon)\sqrt{2 - 2\mathbb{E}_v \mathbb{E}_{v_1, v_2 \in P(v)} v_1^T v_2}$$
$$= (1-\sigma) + \eta(\epsilon)\sqrt{2 + 2L_1}$$

Lemma 2

If the positive instance pairs are divided into 2 parts:

1. In the same subclass, and in the same superclass （O_1）
2. Not in the same subclass, but in the same superclass（O_2）

L_1 has an upper bound which is mainly controlled by O_2

$$\mathbb{E}_v \mathbb{E}_{v_1, v_2 \in P(v)} v_1^T v_2 \geqslant C_\varphi + \frac{(1-\rho)K}{M_1 M_2} s(a_1, a_2)$$

Theorem

Generalization error of G_f can be bounded by the similarity of attention vectors from O_2

$$Err(G_f) \leqslant (1-\sigma) +$$
$$\sqrt{2}\eta(\epsilon)\sqrt{1 - C_\varphi - \frac{(1-\rho)K}{M_1 M_2} \mathbb{E}_{v_1, v_2 \in O_2} s(a_1, a_2)}$$

During the training process, the sample pairs in O_2 can naturally have high similarity attention vectors, thereby reducing the upper bound of the generalization error.

Conclusion

1. We propose the under-study but realistic problem, superclass learning.

2. We propose a novel representation enhancement method (SCLRE) to address superclass learning.

3. We perform extensive experiments to demonstrate that SCLRE outperforms other SOTA classification techniques.

Thanks for Watching!