Superclass Learning with Representation Enhancement

Zeyu Gan[1,2] Suyun Zhao[1,2,*] Jinlong Kang[2] Liyuan Shang[2]
Hong Chen[1,2] Cuiping Li[1,2]
[1]Key Lab of Data Engineering and Knowledge Engineering of MOE Renmin University of China
[2]Renmin University of China, Beijing, China
{zygan, zhaosuyun, kangjinlong, shangliyuan4032, chong, licuiping}@ruc.edu.cn

2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-0129-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/CVPR52729.2023.02304

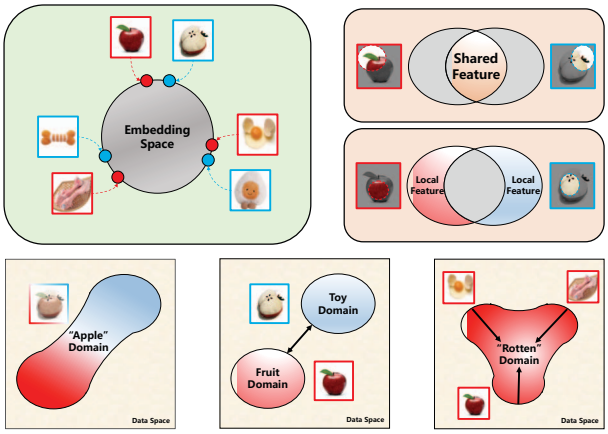
Abstract

In many real scenarios, data are often divided into a handful of artificial super categories in terms of expert knowledge rather than the representations of images. Concretely, a superclass may contain massive and various raw categories, such as refuse sorting. Due to the lack of common semantic features, the existing classification techniques are intractable to recognize superclass without raw class labels, thus they suffer severe performance damage or require huge annotation costs. To narrow this gap, this paper proposes a superclass learning framework, called SuperClass Learning with Representation Enhancement(SCLRE), to recognize super categories by leveraging enhanced representation. Specifically, by exploiting the self-attention technique across the batch, SCLRE collapses the boundaries of those raw categories and enhances the representation of each superclass. On the enhanced representation space, a superclass-aware decision boundary is then reconstructed. Theoretically, we prove that by leveraging attention techniques the generalization error of SCLRE can be bounded under superclass scenarios. Experimentally, extensive results demonstrate that SCLRE outperforms the baseline and other contrastive-based methods on CIFAR-100 datasets and four high-resolution datasets.

1. Introduction

In recent decades, basic-level raw categorization (e.g. cats vs dogs, apples vs bananas) has greatly developed [9, 27] while high-level or super-coarse-grained visual categorization (e.g., recyclable waste vs kitchen waste, creatures vs non-creatures) has received little attention. In many real scenarios, there often exist a handful of high-level categories, wherein numerous images from diverse basic-level categories share one common label. We tend to define this kind of super-coarse-grained class as Superclass.

*Corresponding Author



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

Figure 1. **Illustration of superclass learning.** The samples from a same superclass will be scatteredly distributed in the embedding space. The process of superclass learning is to break old domains and construct new domains. Red indicates the superclass of kitchen waste, and blue indicates the superclass of recyclable waste.

Refuse sorting, as an example, is such a recognization problem with four superclasses, i.e., kitchen waste, recyclable waste, hazardous waste, and others waste. One task of refuse sorting is to accurately collect various items, such as rotten fruits, bones, raw vegetables, and eggshells, into kitchen waste. Traditional recognition needs to identify what exact basic-level categories they are, then sort them out. Obviously, it is wasteful and unrealistic for superclass identification.

Essentially, high-level superclasses contain two characteristics, remarkably distinct from basic-level classes. First, the basic-level classes contained in superclass problems are usually scattered and share few common features. As depicted in the top-left corner subgraph of Fig. 1, the fruit apple, bone, and eggs are remote from each other in feature spaces, though all of them belong to kitchen waste.

979-8-3503-0129-8/23/$31.00 ©2023 IEEE
DOI 10.1109/CVPR52729.2023.02304

Second, the instances from two distinct superclasses may share common features. Just as illustrated in Fig. 1, fruit apple, from kitchen waste, and toy apple, from recyclable waste, are close to each other as they share common semantic features. Obviously, the above-mentioned characteristics indicate that the smoothness assumption [27] under basic-level classification (nearby images tend to have the same label) does not hold in the superclass scenarios. Thus, the existing classification techniques based on smoothness assumption are not practically deployable and scalable and they may suffer severe performance damage in superclass settings. Accordingly, it is valuable and promising to investigate superclass identification.

To tackle the superclass problems, we have to address two main challenges. First, we need to break the original decision boundaries of basic-level classes and disclose a superclass-aware boundary at the basic class level. As depicted in bottom subgraph (a) of Fig. 1, the boundary of the apple domain is original, however, it is useless and even harmful for the refuse sorting as both fruit apples and toy apples belong to this domain. To get the required domain boundary, the apple domain needs to be separated into the fruit domain and toy domain by leveraging their individual local features, as depicted in the bottom subgraph (b) of Fig. 1. In superclass scenarios, it is not enough to investigate the boundary at a basic class level. Consequently, the second challenge is to reconstruct a decision boundary at a superclass level. To achieve this end, it is necessary to merge the domain of classes, such as fruit apples, eggs, and bones into a new rotten superclass domain, as depicted in the bottom subgraph (c) of Fig. 1.

In this paper, we propose a **S**uper**C**lass **L**earning framework with **R**epresentation **E**nhancement. Considering that the semantic representation at the basic class level is not workable for superclass recognition, we propose one cross-instance attention module which could seize the representation across the instances with the same superclass label. By leveraging contrastive adjustment loss, the attention mechanism enhances this representation. Moreover, to overcome the imbalance distribution of superclasses, we adopt target adjustment loss to reconstruct a superclass-aware decision boundary on the enhanced representation space.

In summary, this paper makes the following contributions:

- We propose an under-study but realistic problem, superclass identification, that has notably distinct distribution from basic-level categorization.
- We propose a novel representation enhancement method by leveraging cross-instance attention and then exploit it in superclass identification. And by theoretical analyses, we verify that this self-attention technique can bound the generalization error of superclass recognition.
- Extensive experiments demonstrate that SCLRE outperforms the SOTA classification techniques on one artificial superclass dataset and three real datasets.

The remainder of the paper is organized as follows: in Sec. 2 we briefly review related work and in Sec. 3 we describe our method for superclass recognition. Then, extensive experiments and generalization error analysis are conducted in Secs. 4 and 5. Finally, Sec. 6 draws a brief conclusion.

2. Related Work

Contrastive Learning. The contrastive framework has achieved great success on instance-level problems. Chen et al [3] designed a simple framework to get visual representations in a contrastive way. He et al [14] developed a dynamic dictionary with queues and moving averages to improve contrastive encoder quality. To learn basic-level features and discover more semantic structures for the embedding space, Li et al [22] bridge contrastive learning with clustering. More recently, Caron et al [2] presented a contrastive framework without the need for pairwise comparison computation. To get rid of the limitation of negative samples, Grill et al [12] use two neural networks and predict the representation of the same image under a different augmented view. Chen et al [4] further designed a stop-gradient operation to prevent collapsing. Contrastive learning has also been applied to various research domains and evolved quickly [11, 33, 34]. Instead of focusing on the basic-level boundary, we construct a high-level superclass-aware boundary.

Supervised Contrastive Loss. Traditional contrastive frameworks encode data either unsupervised or self-supervised and apply supervised information in downstream tasks. Khosla et al [19] extend the contrastive mode to a fully-supervised setting and use label information more effectively. To address the impact of long-tailed distribution data, Li et al [23] propose a targeted contrastive loss and guide the encoder to obtain representations based on several preset anchors. Recently, supervised contrastive learning is used in broader research domains both in computer vision and natural language processing [16, 25, 35]. We are inspired by the function of pulling close or away features in these supervised contrastive losses, and therefore propose a new loss, which will pull the features by the superclass labels for better performance in classification.

Self-Attention Module. Self-attention module has been widely used in deep learning. It was first mentioned by Vaswani et al [29] as a part of the transformer model and is widely used in the natural language processing field [5, 21, 24, 26]. Recently, the attention module and transformer also obtained great success in the computer vision

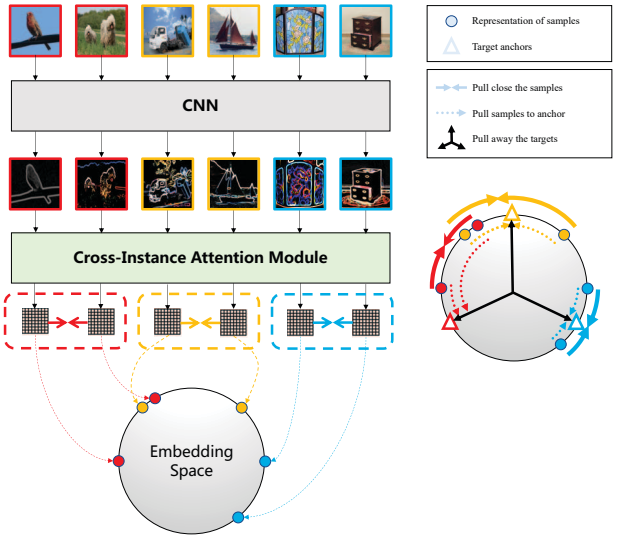


Figure 2. **Illustration of SCLRE.** The images first generate their representations through a convolutional neural network, then mix with each other in a trainable cross-instance attention module for enhancement. After enhancement, the representations are then adjusted according to their superclass labels and the target anchors.

field. Dosovitskiy et al [6] organize image patches as a sequence and propose a vision transformer. Furthermore, He et al [13] propose a masked auto-encoder based on vision transformers. Self-attention modules can also be integrated into convolutional neural networks to improve performance [1, 18, 31]. Fu et al [10] address the fine-grained image recognition problem by using an attention network to find the unique and pivotal features in samples(for example, one kind of bird's tail color). Different from focusing the attention relationship in one sample, we propose a cross-instance attention(CIA) module that cares more about the relationship between instances.

3. Method

Superclass Learning with Representation Enhancement(SCLRE) framework is proposed for superclass image recognition. It exploits the self-attention technique across instances to perform a representation enhancement, thus it achieves the goal to break the basic-level boundary in representations. Then SCLRE reconstructs a new superclass-aware boundary in the enhanced representation space by a series of adjustment losses.

Fig. 2 shows the overview of SCLRE. Sec. 3.1 presents the process of breaking basic-level boundary , while Sec. 3.2 describes the details of constructing the new high-level superclass-aware boundary.

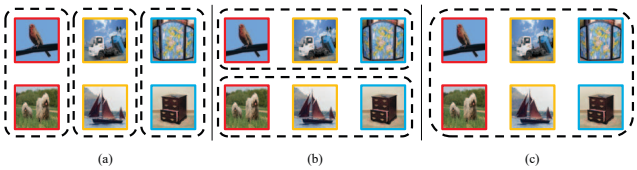


Figure 3. **Different segmentation strategies of cross-instance attention module.** We experiment on three kinds of segmentation strategies and choose the best for the experiment.

3.1. Representation Enhancement

To break the basic-level boundary of representations and obtain valuable enhancements, we mix the representations by leveraging a cross-instance attention module. Unlike the existing attention techniques [32], the cross-instance attention module cares more about the relationship between representations of the instances, rather than the feature representation inside the instance.

Enhanced Representation. Let $\mathcal{X}$ be the data space, $\mathcal{R}^D$ be the D-dimensional embedding space, an encoder $f : \mathcal{X} \to \mathcal{R}^D$ be a mapping from the data space to the embedding space, and an enhancement process $\mathrm{EnH} : \mathcal{R}^D \to \mathcal{R}^D$ be a mapping from the embedding space to the embedding space. For each $x_j \in \mathcal{X}$, the representation $z_j = f(x_j)$, and $Z \in \mathcal{R}^{n \times D}$ is the representation matrix for the batch with size n. The enhanced representation v_j is an output of EnH. More specifically, v_j is defined as follows:

$$v_j = \mathrm{EnH}(z_j; Z) = a_j \times Z, \tag{1}$$

where a_j is a trainable vector for representation z_j computed by cross-instance attention module. Actually, the vector a_j is the j-th row of the attention distribution matrix(ADM) computed in the following Q-K-V mode multi-head attention module [29]:

$$\mathrm{ADM}(Z) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right). \tag{2}$$

The matrix $Q = Z \times W_Q$, where $W_Q \in \mathcal{R}^{D \times d_q}$. Matrix $K = Z \times W_K$, where $W_K \in \mathcal{R}^{D \times d_k}$ and $d_q = d_k$.

The enhancement process breaks the basic-level boundary by merging the semantic features of instances. Cross-instance attention module automatically explores the key instances for representation enhancement and integrates them to obtain a superclass-aware enhanced representation. Moreover, we conduct strict theoretical analysis on the generalization ability of our model in Sec. 5, and discover that the cross-instance module can tightly bound the generalization error of SCLRE.

Data Segmentation Strategy. In the cross-instance attention module, the segmentation strategy of input data is important for representation enhancement. For each instance, the segmentation strategy determines whose repre-

sentations will be received as an enhancement. We considered three kinds of segmentation strategies for the input data in the cross-instance attention module. Fig. 3 illustrates the details of the strategies. In (a), images are gathered in terms of their superclass label, we put the same label images into one batch. In (b), we randomly and evenly select images from the different superclasses and gather them into the batch. In (c), we keep the distribution of the dataset and randomly select the samples to form the batch. The result shows that (c) selecting the samples according to the distribution of the dataset benefits the model most, so we fixed this strategy in the following part of this work.

3.2. SCLRE Loss

To reconstruct the high-level superclass-aware boundary of representations, we design a SCLRE loss to adjust the representations after the attention-weighted enhancement. We keep using the letter v in Eq. (1) to demonstrate representations being enhanced by the cross-instance attention module, and we call it *enhanced representation*. After that, we adjust the representations by taking three aspects into consideration.

Category Classification Loss. We first adopt the cross-entropy loss as the base to obtain a category-level adjustment and prevent the model from collapsing:

$$\mathcal{L}_{ce} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{i,k}\log M(v_i)_k, \tag{3}$$

where N demonstrates batch size, y_i demonstrates the superclass label of the i-th sample in one-hot embedding, $M(v_i)$ demonstrates the output prediction of v_i, and the footscript k demonstrates the k-th component of a vector.

Contrastive Adjustment Loss. Besides classification loss, the samples belonging to the same superclass may share few common features. To obtain a representation-level adjustment, we design a contrastive adjustment loss to pull close the samples in the same superclass and pull away samples in distinct superclasses. We directly regard samples from the same superclass as positive pairs and samples from other superclasses as negative pairs. For one sample v_i and one of its positive sample v^+ (i.e. v_i and v^+ belongs to the same superclass), the contrastive adjustment loss is:

$$\ell(v_i, v^+) = -\log \frac{\exp(s(v_i,v^+)/\tau)}{\exp(s(v_i,v^+)/\tau) + \sum\limits_{v^- \notin P(v_i)} \exp(s(v_i,v^-)/\tau)}. \tag{4}$$

By summing up all the losses of positive pairs and for a batch with size N, we can get the total loss as:

$$\mathcal{L}_{ca} = \sum_{i=1}^{N} \frac{1}{||P(v_i)|| - 1} \sum_{v^+ \in P(v_i)\setminus\{v_i\}} \ell(v_i, v^+), \tag{5}$$

where $P(v_i)$ stands for all the other data that share the same superclass label with v_i in the batch, i.e. $P(v_i) = \{v_k | p(v_k) = p(v_i)\}$, p means the superclass label. $s(\cdot,\cdot)$ stands for the similarity between two vectors, we use cosine similarity for measurement, i.e. $s(v_s, v_t) = \frac{v_s^T v_t}{||v_s|| \cdot ||v_t||}$. τ is the temperature hyperparameter. $||P(v_i)||$ stands for the size of set $P(v_i)$, and it aims to normalize the total loss.

Targeted Adjustment Loss. Since superclass learning naturally faces the problem of lack of classification centers due to the scattered distribution of representations, then follows damage to the constructing process when the contrastive adjustment loss works. To address this problem, we design a targeted adjustment loss to give each superclass a pre-defined category center(i.e. target anchors). We first generate the target anchors and make them far away from each other inspired by [23]. Then we pull every sample close to its pre-defined target anchor by minimizing the following targeted adjustment loss:

$$\mathcal{L}_{ta} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(s(v_i, t(v_i))/\tau)}{\sum_{t\in T}\exp(s(v_i,t)/\tau)}, \tag{6}$$

where N stands for batch size, $s(\cdot,\cdot)$ stands for cosine similarity function, T is the set of all the target anchors, and $t(v_i)$ demonstrates the target anchor allocated to v_i.

SCLRE Adjustment Loss. To construct the high-level superclass-aware boundary with both category-level and instance-level adjustments and overcome the problem of lack of classification center, the total loss of SCLRE is a weighted summation of the above three losses(i.e. Eqs. (3), (5) and (6)):

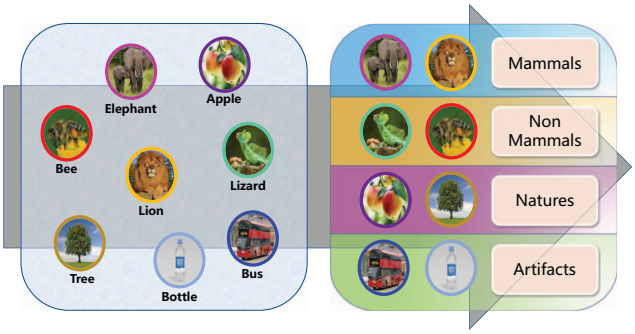
$$\mathcal{L}_{SCLRE} = (1-\alpha)\mathcal{L}_{ce} + \alpha\mathcal{L}_{ca} + \lambda\mathcal{L}_{ta}, \tag{7}$$

where α and λ range from 0 to 1. $\mathcal{L}_{ce}$ and $\mathcal{L}_{ca}$ are designed as a combination to help the encoder explore the concept of superclasses, a larger α means a stronger adjustment strategy and weaker classification ability. $\mathcal{L}_{ta}$ is designed to help the model overcome the imbalanced distribution, and a larger λ means stronger guidance towards the destination of adjustment.

SCLRE adjustment loss can help the model learn more about the concept of the superclass and the high-level categorization boundary. By balancing the influence of cross-entropy loss and contrastive adjustment loss, the model can adjust representations properly. By controlling the utilization of targeted adjustment loss, the model can run smoothly and converge steadily.

4. Experiment

In this section, we conduct extensive experiments on several artificially constructed superclass datasets to demonstrate the effectiveness of our proposed approach, SCLRE.



(a) Original Dataset (b) Superclass Dataset

Figure 4. **Illustration of superclass construction.** We integrate the initial class labels of the original dataset to superclass labels, according to field knowledge. The constructed dataset contains a handful of superclasses.

Dataset	Method	Accuracy(%)
CIFAR100-3	Baseline	72.8
	SupCon [19]	78.1
	SimCLR [3]	79.0
	SCLRE	**80.1**
CIFAR100-4	Baseline	76.0
	SupCon [19]	80.1
	SimCLR [3]	80.6
	SCLRE	**84.0**
CIFAR100-7	Baseline	68.9
	SupCon [19]	72.7
	SimCLR [3]	73.9
	SCLRE	**78.1**

Table 1. **Classification accuracy on low pixels dataset: CIFAR-100.** We compared the classification accuracy on CIFAR100-3, CIFAR100-4, and CIFAR100-7 datasets.

4.1. Experiment Setup

Datasets. In this study, superclass datasets are artificially reorganized from three benchmarks, CIFAR-100, mini-ImageNet [30] and VOC [8], and two real-world datasets, FMoW [20] and Adience [7]. CIFAR-100 contains 50000 training images and 10000 test images from 100 categories with a size of 32 × 32 × 3. mini-ImageNet [30] contains 60000 images in total with a size of 84 × 84 × 3, which are evenly distributed among 100 classes. VOC [8] contains 2501 training images, 2510 validation images and 4952 test images from 20 categories. The real-world dataset FMoW [20] is about a hybrid domain generalization and subpopulation shift problem, which consists of 76863 training images, 11483 validation images and 11327 test images with a size of 224 × 224 × 3. And the real-world dataset Adience [7] contains 26,580 images from 2,284 people.

In our experiments, superclass datasets are artificially constructed by integrating the initial classes, just as illustrated in Fig. 4. To be concrete, We reorganize the 100 classes of CIFAR-100 into 3 superclasses, 4 superclasses, and 7 superclasses, respectively, forming three distinct superclass datasets, CIFAR100-3, CIFAR100-4, and CIFAR100-7. The superclasses in the three data sets are based on movement mode, life form, and are more complex than the original 20 coarse classes in CIFAR-100. For mini-Imagenet and VOC, we reorganize their subclasses into 2 superclasses based on life form. For FMoW, we drop the original labels and turn to predict the geographical location of the images. For Adience, we utilize the age groups as the superclass labels. Thus, those 7 superclass datasets can well simulate the distribution of real-world superclass problems. More reorganization details are listed in the Appendix.

Compared Methods. We compare SCLRE with the baseline model, i.e., ResNet [15], and two SOTA contrastive techniques SupCon [19], and SimCLR [3]. As contrastive techniques perform excellently in representation learning. We adopt the default optimal settings in the training details.

SupCon: In our experiments, SupCon adopts ResNet-50 as backbone, and SupCon loss as the super-classification loss.

SimCLR: It is a known self-supervised contrastive learning framework. In superclass recognization, we keep using ResNet-50 as the backbone and use a supervised downstream task to keep the instance-level features pure, wherein the upstream instance-level results are a kind of learning of the basic level class labels.

Implementation details. We adopt ResNet-50 [15] as our default backbone and we reduce the feature dimension from 2048 to 128 with an extra Multi-Layer Perceptron(MLP) projection head. We train the network using the SGD with a momentum of 0.9 and a batch size of 64. We set the learning rate as 0.001 and the training stage as 200 epochs. Additionally, we resize all the images to 32×32 pixels. For our SCLRE method, we set the key K = 256, value V = 128 and 8 multi-heads are adopted in the attention module.

4.2. Classification Evaluation

Accuracy on CIFAR-100 Datasets. Tab. 1 shows the classification accuracies of SCLRE and the compared methods on CIFAR100-3, CIFAR100-4, and CIFAR100-7, respectively. And we have the following observations: 1) It is observed that SCLRE has a great improvement compared with the baseline. This shows that the cross-instance attention module can effectively activate the contrastive adjustment loss. 2) SCLRE has higher classification accuracies than SimCLR on superclass datasets. This shows that the instance-level semantic information learned by SimCLR may fail in the superclass problem because the images from the same superclass may have quite different, even con-

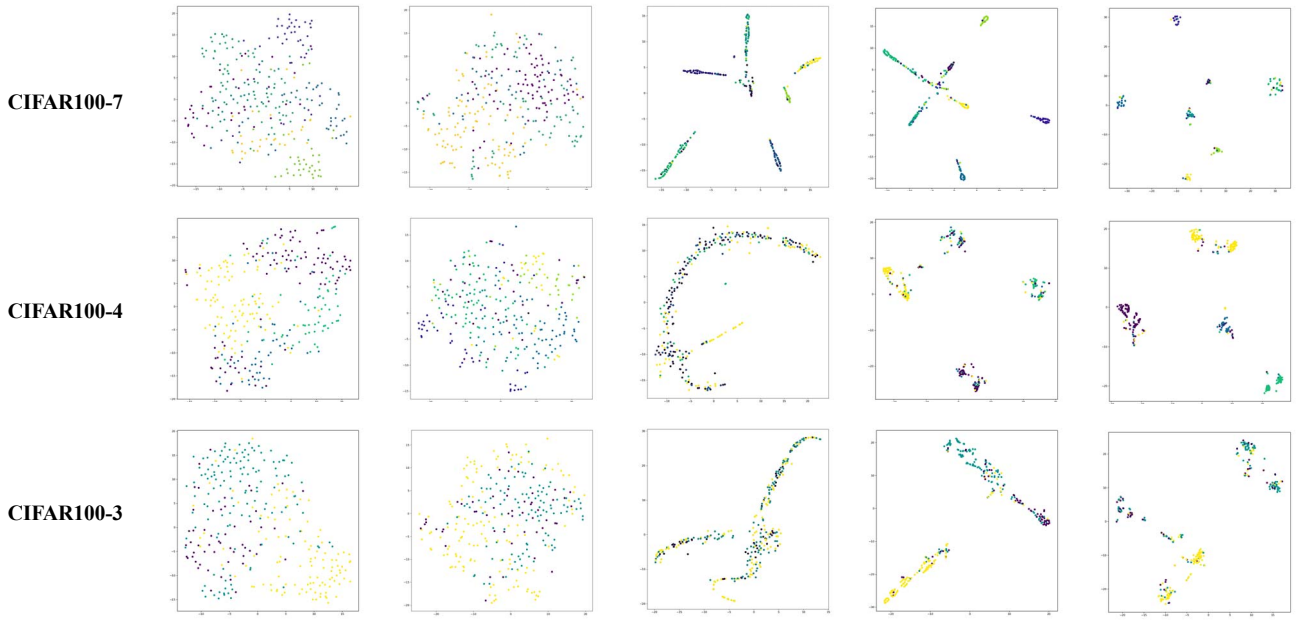


Figure 5. **Visualization results by t-SNE.** We sampled part of the test set and encode them into representations, then visualize them in the embedding space by t-SNE. (a) We directly generate the representations using the backbone as a baseline. (b) We use the upstream encoder of SimCLR. (c) We use the encoder of SupCon. (d) We use part of SCLRE by removing the targeted adjustment loss. (e) We use the entire SCLRE.

tradictory semantic information. Comparatively, SCLRE works well by the superclass-aware decision boundary.

Accuracy on High-Resolution Datasets. Tab. 2 shows the evaluation results on three high-resolution datasets, from which we observe the following facts. 1) we observe that, on the real-world dataset FMoW, SCLRE outperforms other methods by a large margin. On FMoW, the geographic location of building images is adopted as the superclass label, rather than their type. The methods without representation enhancement may perform worse due to the lack of superclass-aware common features. This further confirms the importance of representation enhancement in the superclass problems. 2) SCLRE also outperforms other methods on real-world dataset Adience, which contains 26580 face images from 2284 people. In case every single person is a raw class, the results confirm that SCLRE is also workable in large-scale situations. 3) Moreover, we observe that on FMoW and Adience, SimCLR performs even worse than the baseline. Because the instance-level representation by SimCLR is conflicted with the given superclass labels. 4) The performance on mini-ImageNet appears to be less attractive.This may be because the mini-ImageNet dataset is divided into 2 distinct superclasses: animals and non-animals. For such kind of distribution, basic-level methods

	mini-ImageNet [30]	**FMoW** [20]	**Adience** [7]	**VOC** [8]
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.

can also easily sort them out by exploring the obvious common features inside each superclass(e.g. whether the image contains eyes), which makes it more like a traditional instance-level problem, and SimCLR may perform comparably or even slightly better than SCLRE.

4.3. Visualization

Here we generate the visualization results on representation space by the t-SNE technique [28]. The CIFAR100-3, CIFAR100-4, and CIFAR100-7 datasets are adopted for the visualization experiments.

We observe the following facts from Fig. 5. 1) It is observed that in columns (a) and (b) the images are scattered in their respective representation space, although the images (b) cluster slightly closer than the ones in (a). This further shows that the instance-level representation of SimCLR is

little related to superclass labels. Consequently, in superclass scenarios, SimCLR performs a little better than baseline but worse than SCLRE, just as illustrated in Tab. 1. 2) We observed that the images in column (c) have a more obvious cluster structure than in columns (a) and (b). This shows that the use of superclass labels is important for constructing clear boundaries. But due to the diverse and complex superclass contents, the representations are mostly inaccurate and thus result in less favorable results as reported in Tab. 1. 3) It is also observed that the images in columns (d) and (e) are notably closer than the ones in (a), (b), and (c). This shows that the enhanced representation of SCLRE is superclass-aware. 4) Furtherly, we observe that the images in column (e) are obviously closer than the ones in column (d). This shows that the targeted adjustment loss, one important component of SCLRE, can furtherly cluster images in the same superclasses, accordingly, it benefits reconstructing a superclass-aware decision boundary.

4.4. Sensitivity Analysis

We discuss the hyperparameters α and λ of SCLRE by evaluating their sensitivity on the CIFAR100-7 dataset. When analyzing α, to avoid the influence of target anchors, we removed targeted adjustment loss by fixing λ to 0. We test the sensitivity of α by ranging it from 0 to 1. When analyzing λ, we fix α to 0.5 to avoid its influence.

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

Table 3. **Sensitivity analysis of α.** We range α from 0 to 1 and calculate the accuracy of SCLRE respectively. When $\alpha = 1$, the cross-entropy loss is invalid and the model will lose its classification ability.

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

Table 4. **Sensitivity analysis of λ.** We range λ from 0 to 1 and calculate the accuracy of SCLRE respectively.

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

Table 5. **Sensitivity analysis of batch size.** We vary batch size from 64 to 512 and calculate the accuracy of SCLRE respectively.

Sensetivity of α. Tab. 3 shows the sensitivity of α. Besides extreme situations, we find that α has a little influence on the accuracy. The optimal value of α falls around 0.5 and we fix it in our experiments.

In the case of $\alpha = 0$, the contrastive adjustment loss is invalid and the model degenerates into the baseline. In the case of $\alpha = 1$, the cross-entropy loss is invalid and then the model is not workable for classification.

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**

Table 6. **Ablation study on SCLRE.** We calculated the model size (Params.), computational complexity (FLOPs) and accuracy contributions of four components of SCLRE.

Sensetivity of λ. Tab. 4 demonstrates the sensitivity analysis of λ. λ is a hyperparameter that controls the influence of targeted adjustment loss.

Unlike α, there are two most proper λ values which fall around 0.2 and 0.8. When λ is too small or too large, the performance is both less satisfactory. As a too large λ may be harmful to the accuracy of the model, a too small λ may be short of guidance and has limited performance improvement.

Sensetivity of Batch Size. Intuitively, a larger batch size can bring more abundant samples of superclasses, thus making it easier to construct superclass-aware representations. The experimental results are conducted to verify this statement. Tab. 5 illustrated the results of the sensitivity study of the batch size. The results show that a larger batch size can improve the performance of SCLRE, though the improvement is not significant.

4.5. Ablation Study

Here ablation studies are conducted to demonstrate the contribution of each component in our proposed SCLRE. Just as designed in Sec. 3, SCLRE is composed of the baseline model, contrastive adjustment loss ($\mathcal{L}_{ca}$), cross-instance attention module(CIA), and targeted adjustment loss($\mathcal{L}_{ta}$). The baseline model is the same as our SCLRE without $\mathcal{L}_{ca}$, $\mathcal{L}_{ta}$ and CIA.

According to the results reported in Tab. 6, we have the following observations:

1) It is observed that only CIA itself has an observable contribution to the performance improvements, but CIA joint with $\mathcal{L}_{ca}$ brings notable improvements. This indicates that the CIA module is workable to break the basic class boundary with the great assistance of $\mathcal{L}_{ca}$. They enhance the representation of the superclass by cross-instance attention mechanism. 2) With CIA, $\mathcal{L}_{ca}$ and $\mathcal{L}_{ta}$, SCLRE achieves the optimal improvements. It shows that the targeted adjustment loss does function on reconstructing superclass-aware boundary on the enhanced representation space.

	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**

Table 7. **Robustness research of SCLRE on different backbones.** We calculated the accuracy on a CIFAR100-3 superclass dataset of baseline and SCLRE. We also calculated the improvement SCLRE gets in the different backbones.

4.6. Robustness for Different Backbones

We investigate the robustness of SCLRE by changing the backbone to other convolutional neural networks. Tab. 7 shows the result of the experiments on robustness. The experiments are performed on a CIFAR100-3 superclass dataset and we vary the backbone from smaller networks to larger networks. The result shows that SCLRE can make stable improvements on different backbones.

5. Analysis of Generalization Ability

We conduct an analysis of the generalization ability of SCLRE and prove that the cross-instance attention module can further compress the upper bound of the generalization error. The details of the proof are in the Appendix.

Based on the contrastive adjustment loss in Sec. 3.2, and with ignoring the constant items, we can rewrite the contrastive loss into an expectation form:

$$\begin{aligned}\mathcal{L}_{ca} &\propto -\mathbb{E}_{v,v'}\mathbb{E}_{\substack{v_1,v_2\in P(v)\\ v^-\in P(v')}}\log\frac{\exp(v_1^T v_2)}{\exp(v_1^T v_2)+\exp(v_1^T v^-)}\\ &=\underbrace{-\mathbb{E}_v\mathbb{E}_{v_1,v_2\in P(v)}v_1^T v_2}_{L_1}\\ &+\underbrace{\mathbb{E}_{v,v'}\mathbb{E}_{\substack{v_1,v_2\in P(v)\\ v^-\in P(v')}}\log\left(\exp(v_1^T v_2)+\exp(v_1^T v^-)\right)}_{L_2}\end{aligned}$$

where L_1 measures the alignment between two transformed features and L_2 is the regularizer preventing the collapse of representation.

Lemma 1. As a conclusion in [17], if L_2 is trained to be small than the threshold, which is easily satisfied, the generalization error of downstream classifier G_f has an upper bound:

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

Lemma 2. Sample pairs with the same coarse label consist of two parts: those share the same fine label (O_1) and those have different fine labels (O_2). The former can be closely aligned after contrastive learning. We focus on the latter and prove that:

$$\mathbb{E}_v\mathbb{E}_{v_1,v_2\in P(v)}v_1^T v_2 \geqslant C_\varphi+\frac{(1-\rho)K}{M_1M_2}s(a_1,a_2)$$

where C_φ is a constant determined by the data distribution. M_1, M_2, K are constants. ρ is the ratio of sample pairs with same fine label in all positive pairs.

Theorem 1. The generalization error has another upper bound based on attention vectors according to lemma 1 and lemma 2:

$$\begin{aligned}&Err(G_f)\leqslant(1-\sigma)+\\ &\sqrt{2}\eta(\epsilon)\sqrt{1-C_\varphi-\frac{(1-\rho)K}{M_1M_2}\mathbb{E}_{v_1,v_2\in O_2}s(a_1,a_2))}\end{aligned}$$

The attention vector is a representation of the importance between samples. When we define a superclass, the representation vector z with the same coarse label will be pulled together by the contrastive loss. During this process, samples in the superclass will focus on the same important enhanced representations, thus leading to similar vectors and declining the upper bound of generalization error.

6. Conclusion

In this study, we explore the high-level categories recognition problem, i.e., superclass categorization. And a superclass learning framework, that exploits self-attention techniques cross the instances to enhance the representation, is proposed. Thus, the distribution of superclass is modified in the enhanced representation space and a new superclass-aware decision boundary is then reconstructed. In theory, the generalization error of SCLRE can be bounded by attention constraints. In extensive experiments, SCLRE outperforms the SOTA classification methods.

In the near future, we would like to tackle the superclass problems under the scenarios without full annotation by leveraging semi-supervised learning and active learning.

Acknowledgements. This work was supported by the National Key Research & Develop Plan(2018YFB1004401); National Natural Science Foundation of China (62276270,62072460, 62172424); Beijing Natural Science Foundation (4212022). it is also partially supported by the Opening Fund of Hebei Key Laboratory of Machine Learning and Computational Intelligence.

References

[1] Amjad Almahairi, Nicolas Ballas, Tim Cooijmans, Yin Zheng, H. Larochelle, and Aaron C. Courville. Dynamic capacity networks. In *ICML*, 2016. 3

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020. 2

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *ArXiv*, abs/2002.05709, 2020. 2, 5

[4] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15745–15753, 2021. 2

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 2

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2021. 3

[7] Eran Eidinger, Roee Enbar, and Tal Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on information forensics and security*, 9(12):2170–2179, 2014. 5, 6

[8] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 2009. 5, 6

[9] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010. 1

[10] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4476–4484, 2017. 3

[11] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *ArXiv*, abs/2104.08821, 2021. 2

[12] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 2

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 3

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020. 2

[15] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[16] Qiushi Huang, Tom Ko, H Lilian Tang, Xubo Liu, and Boyong Wu. Token-level supervised contrastive learning for punctuation restoration. In *Interspeech*, 2021. 2

[17] Weiran Huang, Mingyang Yi, and Xuyang Zhao. Towards the generalization of contrastive self-supervised learning. *arXiv preprint arXiv:2111.00743*, 2021. 8

[18] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *NIPS*, 2015. 3

[19] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *ArXiv*, abs/2004.11362, 2020. 2, 5

[20] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 2021. 5, 6

[21] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019. 2

[22] Junnan Li, Pan Zhou, Caiming Xiong, Richard Socher, and Steven C. H. Hoi. Prototypical contrastive learning of unsupervised representations. *ArXiv*, abs/2005.04966, 2021. 2

[23] Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogério Schmidt Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6908–6918, 2022. 2, 4

[24] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019. 2

[25] Zheda Mai, Ruiwen Li, Hyunwoo J. Kim, and Scott Sanner. Supervised contrastive replay: Revisiting the nearest class mean classifier in online class-incremental continual learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3584–3594, 2021. 2

[26] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 2

[27] Lars Schmarje, Monty Santarossa, Simon-Martin Schroder, and Reinhard Koch. A survey on semi-, self- and unsupervised learning for image classification. *IEEE Access*, 9:82146–82168, 2021. 1, 2

[28] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 6

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 3

[30] Oriol Vinyals, Charles Blundell, Timothy P. Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *NIPS*, 2016. 5, 6

[31] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In-So Kweon. Cbam: Convolutional block attention module. In *ECCV*, 2018. 3

[32] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8805–8814, 2020. 3

[33] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *ArXiv*, abs/2010.13902, 2020. 2

[34] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 833–842, 2021. 2

[35] Mingkai Zheng, Fei Wang, Shan You, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Weakly supervised contrastive learning. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10022–10031, 2021. 2