OOD-CV Challenge Report

September 18, 2023

1 Team details

- Challenge track: GCD Track
- Team name: Align
- Team leader name: Chuyu Zhang
- Team leader address, phone number, and email: 393 Middle Huaxia Road Pudong, Shanghai, 201210, China. +86 15072762954. zhangchy2@shanghaitech.edu.cn
- Rest of the team members: Peiyan Gu, Xuming He
- Team website URL:
- Affiliation: ShanghaiTech University
- User names on the OOD-CV Codalab competitions: kleinwhu
- Link to the codes of the solution(s): https://github.com/kleinzcy/ARA

2 Contribution details

- Title of the contribution:
  A novel adaptive representation alignment learning framework for generalized category discovery

- General method description:
  We propose an adaptive representation alignment learning framework for generalized category discovery. Our core idea is two folds: 1) we first utilize the known classes to establish a pre-trained representation space, which can extract known-class knowledge and avoid the influence of noisy learning from unlabeled data; 2) we adopt the pre-trained known-class model as a prior to guide the discovery of novel classes, in which the representation space of known and novel classes can be adaptively aligned with the pre-trained known-class representation space.

  Specifically, our framework consists of two learning stages. In the first stage, we perform supervised learning on known classes to obtain the pre-trained representation space containing known-class knowledge. In the second stage, we learn a joint model on known and novel classes. To mitigate the noisy learning effect of unlabeled data and guide the learning of novel classes, we align the representation of the joint model with the pre-trained known-class representation. We propose a novel contrastive knowledge distillation term to implement our adaptive representation alignment constraint and develop a negative sample generation strategy based on mixup. More importantly, to adaptively align representations, we introduce an adapter layer to transform the pre-trained known-class representation space to the joint representation space, and an instance-wise mask layer, which selects only part of the feature after the adapter layer for alignment, enabling more flexible learning of novel classes.

- References:
  [1] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive Representation Distillation.
  [2] Sagar Vaze, Kai Han, Andrea Vedaldi, Andrew Zisserman. Generalized Category Discovery
  [3] Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, Julien Mairal, Piotr Bojanowski, Armand Joulin. Emerging Properties in

Self-Supervised Vision Transformers.
[4] Xin Wen, Bingchen Zhao, Xiaojuan Qi. Parametric Classification for Generalized Category Discovery: A Baseline Study.
[5] Zhun Zhong, Linchao Zhu, Zhiming Luo, Shaozi Li, Yi Yang, Nicu Sebe. OpenMix: Reviving Known Knowledge for Discovering Novel Visual Categories in An Open World.

- Representative image / diagram of the method(s):

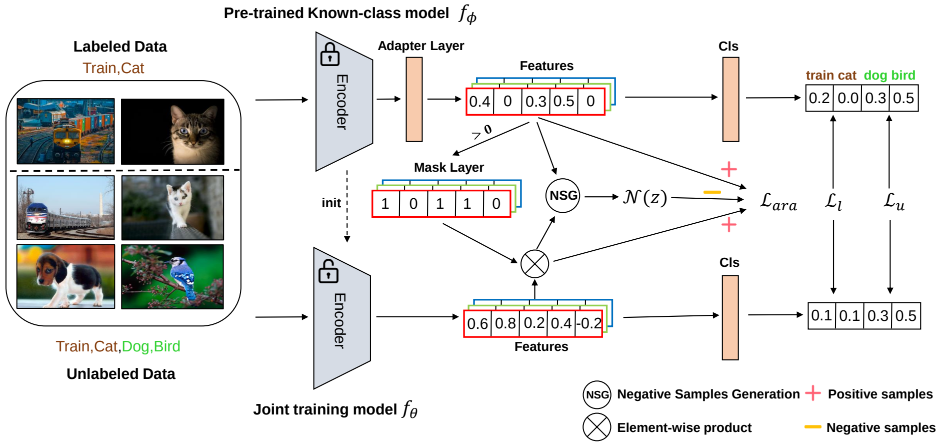

Figure 1: The overview of our adaptive representation alignment framework.

3 Global Method Description

- Total method complexity:

- Model Parameters: 85.95M

- Run Time: 15min for pre-train known-class model training+ 1.5h for adapter layer training + 2h for joint model training

- Which pre-trained or external methods / models have been used: We use the pre-trained known-class model.

- Training description:
  We adopt a two-stage learning strategy to learn our adaptive representation alignment framework. The first stage involves training the model (f_ϕ) on labeled known class data using the standard cross-entropy loss to extract the knowledge of know classes. In the second stage, as shown in Fig.1, we first learn an adapter layer, which we will illustrate in Sec. 3.2, and then utilize the pre-trained known-class model (f_ϕ) to guide the learning of joint representation space (f_θ). To learn the joint representation space (f_θ) and cosine classifier (h), we introduce a comprehensive loss function consisting of three components. The first component is supervised loss, which focuses on known class data and facilitates knowledge extraction from these classes. The second component is the self-labeling loss applied to unlabeled data, aiming to classify known classes and cluster novel classes. The third component is our proposed adaptive representation alignment constraint, which can mitigate the noisy learning effect of unlabeled data and guide joint representation learning. We propose a novel contrastive loss to implement the constraint, which treats the representations of the same unlabeled data in two representation spaces as positive pairs while considering all other unlabeled data as negative samples. In summary, the loss function of the joint representation space can be written as:

  $$\mathcal{L} = (1-\alpha)\mathcal{L}_s + \alpha\mathcal{L}_u + \beta\mathcal{L}_{ara} \tag{1}$$

  where $\mathcal{L}_s$ is the typical cross-entropy loss on labeled known-class data, $\mathcal{L}_u$ is the self-labeling loss on unlabeled data, and $\mathcal{L}_{ara}$ is the adaptive representation alignment regularization term for unlabeled data. α, β are hyperparameters that control the weight of loss. In the following, we will first revisit the $\mathcal{L}_u$, and then provide a detailed explanation of our novel $\mathcal{L}_{ara}$.

3.1 Revisiting Self-labeling Loss $\mathcal{L}_u$

In this section, for completeness, we revisit the details of self-labeling loss[3], which learns clustering and representations concurrently for unlabeled data. Specifically, for each unlabeled data point x_i, we generate two views $x_i^{v_1}$ and $x_i^{v_2}$ through random data augmentation. These views are then fed into the ViT encoder and cosine classifier

(h), resulting in two predictions $\mathbf{y}_i^{v_1} = h(f_\theta(x_i^{v_1}))$ and $\mathbf{y}_i^{v_2} = h(f_\theta(x_i^{v_2}))$, $\mathbf{y}_i^{v_1}, \mathbf{y}_i^{v_2} \in \mathbb{R}^{C^k+C^n}$. As we expect the model to produce consistent predictions for both views, we employ $\mathbf{y}_i^{v_2}$ to generate a pseudo label for supervising $\mathbf{y}_i^{v_1}$. The probability prediction and its pseudo label are denoted as:

$$\mathbf{p}_i^{v_1} = \mathtt{Softmax}(\mathbf{y}_i^{v_1}/\tau), \quad \mathbf{q}_i^{v_2} = \mathtt{Softmax}(\mathbf{y}_i^{v_2}/\tau') \tag{2}$$

Here, τ, τ' represents the temperature coefficients that control the sharpness of the prediction and pseudo label, respectively. Similarly, we employ the generated pseudo-label $\mathbf{q}_i^{v_1}$, based on $\mathbf{y}_i^{v_1}$, to supervise $\mathbf{y}_i^{v_2}$. However, self-labeling approaches may result in a degenerate solution where all novel classes are clustered into a single class. To mitigate this issue, we introduce an additional constraint on cluster size. Thus, the loss function can be defined as follows:

$$\mathcal{L}_u = \frac{1}{2|\mathcal{D}^u|} \sum_{i=1}^{|\mathcal{D}^u|} [l(\mathbf{p}_i^{v_1}, \mathtt{SG}(\mathbf{q}_i^{v_2})) + l(\mathbf{p}_i^{v_2}, \mathtt{SG}(\mathbf{q}_i^{v_1}))] + \epsilon \mathbf{H}(\frac{1}{2|\mathcal{D}^u|} \sum_{i=1}^{|\mathcal{D}^u|} \mathbf{p}_i^{v_1} + \mathbf{p}_i^{v_2}) \tag{3}$$

Here, $l(\mathbf{p}, \mathbf{q}) = -\mathbf{q} \log \mathbf{p}$ represents the standard cross-entropy loss, and $\mathtt{SG}$ denotes the "stop gradient" operation. The entropy regularizer $\mathbf{H}$ enforces cluster size to be uniform thus alleviating the degenerate solution issue. The parameter ϵ represents the weight of the regularizer.

We note that self-labeling loss is not our contribution, and our method does not rely on the design of $\mathcal{L}_u$, and it can be replaced by any other clustering loss.

3.2 Adaptive Representation Alignment Loss $\mathcal{L}_{ara}$

As discussed above, the noisy learning of unlabeled data affects knowledge transfer, while the pre-trained known-class representation encompasses valuable information. To exploit known-class knowledge and facilitate the discovery of novel classes, we propose a novel adaptive representation alignment framework, which utilizes the pre-trained known-class model as a prior to guide the learning of novel classes. Our constraint tends to maintain the structure of pre-trained known-class representation space, thus mitigating the impact of noisy learning from unlabeled data on knowledge transfer. We develop a novel contrastive loss to impose this constraint. And we design a negative sample generation strategy to mitigate any potential adverse effects of contrastive learning on clustering novel classes. Moreover, we introduce an adapter layer and a mask layer to achieve more effective and flexible representation alignment. In the following, we provide a detailed description of each component in our method.

Contrastive Representation Alignment We first present the naive representation alignment constraint implemented by contrastive loss[1]. Specifically, we take the representations of the unlabeled data x_i in two representation spaces, $\mathbf{z}_i^S = f_\phi(x_i)$ and $\mathbf{z}_i = f_\theta(x_i)$, as a positive pair while taking the representations of other data in two representation spaces as negative samples. Therefore, the naive contrastive representation alignment constraint term is formulated as :

$$\mathcal{L}_{cRA} = -\frac{1}{|\mathcal{D}^u|}\sum_{i=1}^{|\mathcal{D}^u|}\log\frac{e^{\mathbf{z}_i^\top \mathbf{z}_i^S/\tau}}{e^{\mathbf{z}_i^\top \mathbf{z}_i^S/\tau}+\sum_{\mathbf{z}\in\mathcal{N}(\mathbf{z})}e^{\mathbf{z}_i^\top \mathbf{z}/\tau}} \tag{4}$$

where $\mathcal{N}(\mathbf{z})$ is the set of the negative samples in memory. Note that we align the representations of all unlabeled data instead of only novel class data. The naive contrastive representation alignment minimizes the distance of two representations, thus maintaining the knowledge of known classes contained in the pre-trained known-class model.

Negative Samples Generation The use of contrastive loss poses a potential issue as it may mistakenly treat different unlabeled data samples from the same class as negative samples. To address this concern, we generate negative samples by combining the representation of labeled and unlabeled data. Specifically, we mix the representations of labeled and unlabeled data as follows:

$$\mathcal{N}(\mathbf{z}) = \{\mathbf{z}|\mathbf{z} = \eta\mathbf{z}^l + (1-\eta)\mathbf{z}^u, \eta \in (0.5, 1]\} \tag{5}$$

Here, $\mathbf{z}^l$ and $\mathbf{z}^u$ represent the representations of labeled and unlabeled data in two representation spaces, and η is a random value between 0.5 and 1. Because $\eta > 0.5$, the generated negative samples tend to be biased towards the known classes since the labeled data belong to known classes. Consequently, this approach helps to avoid class collision issues for novel classes.

Adapter Layer Although the representation space initialized with known classes contains rich semantic information and can represent novel classes well, it has not encountered novel classes. Therefore, directly aligning the two representation spaces would make the joint representation space overly biased towards known classes, which is not conducive to jointly classifying known classes and discovering novel classes. To mitigate this issue, we propose a simple adapter layer f_w that transforms the representation space initialized with known classes to the joint representation space, which is more beneficial for known and novel classes learning. We denote the transformation process as follows:

$$\mathbf{v} = f_w(f_\phi(x)) \tag{6}$$

Specifically, our adapter layer consists of a linear and a ReLU layer. The original representation space is linearly transformed and truncated so that the transformed space can retain most of the original structure. To learn the adapter layer, we utilize $\mathbf{v}$ to perform classification and clustering, and adopt $\mathcal{L}_s, \mathcal{L}_u$ to learn labeled and unlabeled data, respectively. The total loss is denoted as $(1-\alpha)\mathcal{L}_s + \alpha\mathcal{L}_u$, which is the same as first two terms in Eqn.(1).

Mask Layer The proposed representation alignment strategy aligns all features in two representation spaces, imposing a strong constraint on learning joint representation space. To relax this constraint, we propose an instance-wise mask layer that selects some features for alignment and does not impose any constraints on the unselected features. Since our adapter layer contains a ReLU layer, features less than 0 do not contribute to the final classification and clustering. Therefore, we only utilize features greater than 0 after the adapter layer to constraint joint model learning. Consequently, the feature after the mask layer is denoted as:

$$\mathbf{u} = \mathbb{1}(\mathbf{v} > 0) \cdot \mathbf{z} \tag{7}$$

where $\mathbb{1}$ is the indicator function, $\mathbf{v}$ is the representation of the pre-trained known-class model after the adapter layer, and $\mathbf{z}$ is the representation of the joint learning model. For an unlabeled sample, if more features are greater than 0 after the adapter layer, the constraint will be stronger, and vice versa. Consequently, the mask layer enables our representation alignment term adaptive to each instance.

In summary, with the above components, our novel adaptive representation alignment loss can be written as:

$$\mathcal{L}_{ara} = -\frac{1}{|\mathcal{D}^u|}\sum_{i=1}^{|\mathcal{D}^u|}\log\frac{e^{\mathbf{u}_i^\top \mathbf{v}_i/\tau}}{e^{\mathbf{u}_i^\top \mathbf{v}_i/\tau}+\sum_{\mathbf{z}\in\mathcal{N}(\mathbf{z})} e^{\mathbf{u}_i^\top \mathbf{z}/\tau}} \tag{8}$$

where $\mathcal{N}(\mathbf{z}), \mathbf{v}, \mathbf{u}$ are defined in Eqn.(5)(6)(7), respectively. With our novel adaptive representation alignment loss, the joint representation learning can maintain the knowledge of known classes, and mitigate the effect of noisy learning introduced by unlabeled data. Moreover, it preserves the potential relation between known and novel classes in the pre-trained representation space, promoting knowledge transfer between them.

- Testing description:
  We only use the joint model in test time.

- Quantitative and qualitative advantages of the proposed solution:
  See the quantitative advantages of the proposed solution in the above section.

  Qualitative advantages:

Table 1: Ablation study.

cRA	NSG	AL	ML	CUB			Aircraft			Scars		
				All	Known	Novel	All	Known	Novel	All	Known	Novel
				61.7	68.0	58.5	49.6	56.3	46.2	51.8	71.9	42.0
✓				65.0	73.1	61.0	53.5	61.8	49.3	55.8	76.9	45.6
✓	✓			65.5	71.9	62.2	52.9	59.5	49.6	58.0	77.4	48.5
✓	✓	✓		66.8	**75.6**	62.5	55.6	60.5	53.1	57.6	75.9	48.8
✓	✓	✓	✓	**67.1**	73.7	**63.8**	**55.9**	**60.7**	**53.6**	**59.2**	**79.1**	**49.6**

cRA, NSG, AL, and ML denote contrastive representation alignment, negative samples generation, adapter layer, and mask layer respectively.

Note that all the experiments in Tab 1 and Tab 2 were conducted in the same dataset settings of [2] which are slightly different from the OOD-CV Challenge's dataset settings. In competition, we follow the GCD Track settings.

- Results of the comparison to other approaches (if any):

Table 2: Comparison with state-of-the-art methods.

Method	CIFAR100-80			ImageNet100-50			CUB			Scars			Aircraft			Herbarium19		
	All	Known	Novel	All	Known	Novel	All	Known	Novel	All	Known	Novel	All	Known	Novel	All	Known	Novel
K-means	52.0	52.2	50.8	72.7	75.5	71.3	34.3	38.9	32.1	12.8	10.6	13.8	16.0	14.4	16.8	12.9	12.9	12.8
RS+	58.2	77.6	19.3	37.1	61.1	24.8	33.3	51.6	24.2	28.3	61.8	12.1	26.9	36.4	22.2	27.9	55.8	12.8
UNO	69.5	80.6	47.2	70.3	95.0	57.9	35.1	49.0	28.1	35.5	70.5	18.6	40.3	56.4	32.2	28.3	53.7	14.7
ORCA	69.0	77.4	52.0	73.5	92.6	63.9	35.3	45.6	30.2	23.5	50.1	10.7	22.0	31.8	17.1	20.9	30.9	15.5
GCD	70.8	77.6	57.0	74.1	89.8	66.3	51.3	56.6	48.7	39.0	57.6	29.9	45.0	41.1	46.9	35.4	51.0	27.0
PromptCAL	81.2	**84.2**	75.3	83.1	92.7	78.3	62.9	64.4	62.1	50.2	70.1	40.6	52.2	52.2	52.3	-	-	-
Ours	**82.8**	84.0	**80.3**	**84.1**	**92.8**	**79.7**	**67.1**	**73.7**	**63.8**	**59.2**	**79.1**	**49.6**	**55.9**	**60.7**	**53.6**	**43.0**	**56.2**	**35.9**

- Novelty of the solution and if it has been previously published:
  We introduce a novel adaptive representation alignment framework for generalized category discovery aiming at effectively leveraging the knowledge in known classes to discover novel classes. Our framework follows a two-stage approach, starting with the initialization of the representation space using known class data, followed by joint training on both known and novel class data to facilitate the discovery of novel classes. During this joint learning process, we distill knowledge from known classes by aligning the representations of unlabeled data in two distinct representation spaces. To this end, we propose a novel contrastive loss to implement the representation alignment constraint. In addition, we introduce a negative sample generation strategy based on mixup to mitigate any adverse effects of contrastive learning on clustering novel classes. To enhance the efficiency and flexibility of our alignment, we incorporate an adapter layer and a mask layer.

4 Technical details

- Language and implementation details (including platform, memory, parallelization requirements) :
  We use Python. We adopt the DINO[3] pre-trained ViT-B/16 as our backbone, and we only finetune the last block of ViT-B/16. The adapter layer is composed of a linear and ReLU layer. In the first stage, we train our model by 30 epochs on labeled data. In the second stage, we train our model by 100 epochs on all data. We adopt the SGD optimizer with a momentum of 0.9, a weight decay of 5×10^{-5}, and an initial learning rate of 0.1, which reduces to $1e-4$ at 100 epoch using

a cosine annealing schedule. The batch size is 128 and the size of the negative set $\mathcal{N}(\mathbf{z})$ is 2048. The data augmentation is the same as [2]. For hyperparameters, we follow [2] to set $\alpha = 0.35, \epsilon = 1$. Moreover, we follow [3] to set τ to 0.1, and τ' is initialized to 0.07, then warmed up to 0.04 with a cosine schedule in the starting 30 epochs. For the additional hyperparameter β that we introduced, we set it to 0.1 for all datasets. We then validate its sensitivity in the ablation study. All the experiments are conducted on a single NVIDIA TITAN RTX with 24GB.

- Human effort required for implementation, training and validation?:
  No.

- Runtime at test per image:
  4.5ms.

- Comment the efficiency of the proposed solution(s)? :
  We conduct an extensive experiments on six benchmark datasets, and the results demonstrate the superiority of our approach over the previous state-of-the-art methods.