Semantically Coherent Out-of-Distribution Detection

Jingkang Yang[1], Haoqi Wang[2], Litong Feng[2], Xiaopeng Yan[2], Huabin Zheng[2],
Wayne Zhang[2,3,4], Ziwei Liu[1]✉
[1] S-Lab, Nanyang Technological University [2] SenseTime Research
[3] Qing Yuan Research Institute, Shanghai Jiao Tong University
[4] Shanghai AI Laboratory, Shanghai, China
jingkang001@ntu.edu.sg, {lastname.firstname}@sensetime.com, ziwei.liu@ntu.edu.sg

Abstract

Current out-of-distribution (OOD) detection benchmarks are commonly built by defining one dataset as in-distribution (ID) and all others as OOD. However, these benchmarks unfortunately introduce some unwanted and impractical goals, e.g.*, to perfectly distinguish CIFAR dogs from ImageNet dogs, even though they have the same semantics and negligible covariate shifts. These unrealistic goals will result in an extremely narrow range of model capabilities, greatly limiting their use in real applications. To overcome these drawbacks, we re-design the benchmarks and propose the* ***semantically coherent out-of-distribution detection (SC-OOD)****. On the SC-OOD benchmarks, existing methods suffer from large performance degradation, suggesting that they are extremely sensitive to low-level discrepancy between data sources while ignoring their inherent semantics. To develop an effective SC-OOD detection approach, we leverage an external unlabeled set and design a concise framework featured by* ***unsupervised dual grouping (UDG)*** *for the joint modeling of ID and OOD data. The proposed UDG can not only enrich the semantic knowledge of the model by exploiting unlabeled data in an unsupervised manner, but also distinguish ID/OOD samples to enhance ID classification and OOD detection tasks simultaneously. Extensive experiments demonstrate that our approach achieves the state-of-the-art performance on SC-OOD benchmarks. Code and benchmarks are provided on our project page: https://jingkang50.github.io/projects/scood.*

1. Introduction

Although dominated on visual recognition [1, 2], deep learning models are still notorious for the following two drawbacks: **1)** their performance endures a significant drop when test data distribution has a large covariate shift from training [3]; **2)** they tend to recklessly classify a test image into a certain training class, even though it has a semantic shift from training, that is, it may not belong to any training class [4]. Those defects seriously reduce the model trustworthiness and hinder their deployment in real, especially high-risk applications [5, 6]. To solve the problem, out-of-distribution (OOD) detection aims to distinguish and reject test samples with either covariate shifts or semantic shifts or both, so as to prevent models trained on in-distribution (ID) data from producing unreliable predictions [4].

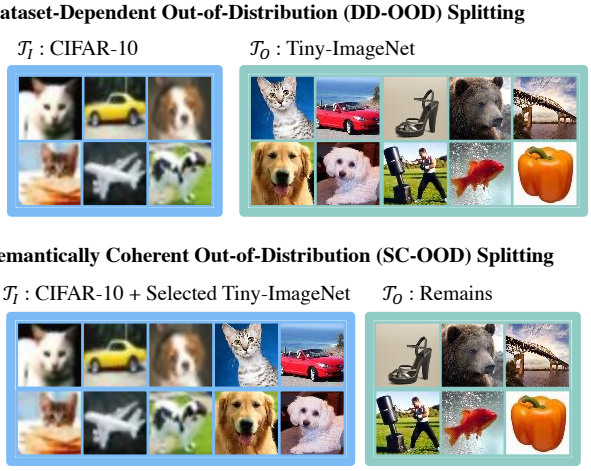


Figure 1: **Dataset-dependent OOD (DD-OOD) *vs*. Semantically Coherent OOD (SC-OOD).** We notice that a certain number of DD-OOD samples have the same semantics as ID, with negligible covariate shifts. Here, we redirect these samples back to ID in the SC-OOD setting.

Existing OOD detection methods mostly focus on calibrating the distribution of the softmax layer [4] through temperature scaling [7], generative models [8, 9], or ensemble methods [10, 11]. Other solutions collect enormous OOD samples to learn the ID/OOD discrepancy dur-

Table 1: **Performance of ODIN [7], MCD [14], and Energy-Based OOD (EBO) [15] on DD-OOD/SC-OOD settings.** Previous methods achieve nearly perfect results on DD-OOD but suffer a drastic drop on SC-OOD. [1]

Method	FPR95 ↓		AUROC ↑		AUPR ↑	
	DD	SC	DD	SC	DD	SC
ODIN (ICLR18)	0.46	55.0	99.8	88.8	99.8	84.2
EBO (NeurIPS20)	1.56	50.6	99.5	90.4	99.4	85.4
MCD (ICCV19)	0.01	68.6	99.9	88.9	99.9	82.1

ing training [12, 13, 14]. Appealing experimental results are achieved by existing methods. For example, MCD [14] reports *near-perfect* scores across the classic benchmarks. The OOD detection problem seems completely solved.

However, by scrutinizing the common-used OOD detection benchmarks [4, 7, 13], we discover some irrationality on OOD splitting. Under the assumption that different datasets represent different data distributions, current benchmarks are commonly built by defining one dataset as ID and all others as OOD. Figure 1-a shows one popular benchmark that uses the entire CIFAR-10 test data as ID and the entire Tiny-ImageNet test data as OOD. However, we observed that around 15% Tiny-ImageNet test samples actually shares the same semantics with CIFAR-10's ID categories (*ref.* Section 4). For example, Tiny-ImageNet contains six dog-breeds (*e.g.* golden retriever, Chihuahua) that match CIFAR-10's class of dog, while their covariate shifts are negligible. In this case, the perfect performance on the above dataset-dependent OOD (DD-OOD) benchmarks may indicate that models are attempting to overfit the low-level discrepancy on the negligible covariate shifts between data sources while ignoring inherent semantics. This fails to meet the requirement of realistic model deployment.

To overcome the drawbacks of the DD-OOD benchmarks, in this work, we re-design **semantically coherent out-of-distribution detection (SC-OOD)** benchmarks, which re-organize ID/OOD set based on semantics and only focus on real images where the covariate shift can be ignored, as depicted in Figure 1-b. In this case, the ID set becomes semantically coherent and different from OOD. Existing methods suffer a large performance degradation on revised SC-OOD benchmarks as shown in Table 1, indicating that the OOD detection problem is still unresolved. [1]

For an effective SC-OOD approach, we leverage an external unlabeled set like OE [13]. *Different from OE [13] whose unlabeled set is purely OOD, our unlabeled set is contaminated by a portion of ID samples.* We believe it is a more realistic setting, as a powerful image crawler can easily prepare millions of unlabeled data but will inevitably introduce ID samples that are expensive to be purified.

With a realistic unlabeled set for SC-OOD, we design an elegant framework featured by **unsupervised dual grouping (UDG)** for the joint modeling of labeled and unlabeled data. The proposed UDG enhances the semantic expression ability of the model by exploring unlabeled data with an unsupervised deep clustering task, and the grouping information generated by the auxiliary task can also dynamically separate the ID and OOD samples in the unlabeled set. ID samples separated from the unlabeled set will join other given ID samples for classifier training, and the rest will be forced to produce a uniform posterior distribution like other OE methods [13]. In this way, ID classification and OOD detection performances are simultaneously improved.

To summarize, the contributions of our paper are: **1)** We highlight the problem of current OOD detection benchmarks and re-design them to address semantic coherency in out-of-distribution detection. **2)** A concise framework using realistic unlabeled data is proposed, featuring the unsupervised dual grouping which not only enriches the semantic knowledge of the model in an unsupervised manner, but also distinguishes ID/OOD samples to enhance ID classification and OOD detection tasks simultaneously. **3)** Extensive experiments demonstrate our approach achieves state-of-the-art performance on SC-OOD benchmarks.

2. Related Works

Out-of-Distribution Detection. OOD detection aims to distinguish test images that come from different distribution comparing to training samples [4]. The simplest baseline uses the maximum softmax probability (MSP) to identify OOD samples, which is based on the observation that DNNs tend to produce lower prediction probability for misclassified and OOD inputs [4]. Follow-up work uses various techniques to improve MSP. ODIN [7] applies temperature scaling on the softmax layer to increase the separation between ID and OOD probabilities. Small perturbations are also introduced into the input space for further improvement. Some probabilistic methods attempt to model the distribution of the training samples and use the likelihood, or density, to identify OOD samples [8, 9, 16]. Besides, ensemble methods can also be used to robustify the models [10, 11]. Recently, energy scores from the energy-based model are found theoretically consistent with the probability density and suitable for OOD detection [15]. All the above methods only rely on ID samples for OOD detection.

Another group of methods for OOD detection utilizes a set of external OOD data, based on which the discrepancy between ID and OOD data is learned. The baseline work for this branch is OE [13]. Based on the MSP baseline [4], a large-scale selected OOD set is introduced as outlier expo-

[1] In Table 1, both DD and SC-OOD benchmarks consider CIFAR-10 as ID and Tiny-ImageNet as OOD. All methods are tested using their released DenseNet models. AUPR corresponds to AUPR-Out in Section 4.

sure (OE) and an additional objective is included, expecting DNNs to produce uniform softmax scores for extra samples. Afterwards, MCD [14] proposes a network with two classifiers, which are forced to produce maximum entropy discrepancy for extra OOD samples. Some works explore the optimal strategies for external OOD data sampling [17]. However, we find two problems with the previous method of using external OOD data: **1)** in the realistic setting, a purified OOD set is difficult to obtain, as ID samples are inevitably introduced and expensive to filter out. **2)** current methods only regard OOD set holistically, neglecting the abundant semantic information within the set. In this paper, we take a realistic unlabeled set that is a natural ID/OOD mixture, and hope to well explore the knowledge within it.

Deep Clustering. Deep clustering is an unsupervised learning method that trains DNNs using the cluster assignments of the resulting features [18]. In this paper, we integrate it into our main proposed unsupervised dual grouping (UDG), aiming to not only learn visual representations through self-supervised training on both labeled and unlabeled sets, but also do cluster-wise OOD probability estimation to filter out ID samples from the unlabeled set described in Section 3.4.

Besides clustering, other unsupervised methods can also be implemented as auxiliary (pretext) tasks, *i.e.*, patch orderings [19, 20], colorization [21], rotation prediction [22] and contrastive learning [23, 24, 25]. Although we believe that they can help discover latent knowledge in the unlabeled set and enhance visual representation, their interactive potential on the primary ID filtering task is relatively limited compared with the clustering auxiliary task.

3. Our Approach

In this section, we introduce our proposed end-to-end pipeline with unsupervised dual grouping (UDG) in detail.

3.1. Problem Statement

Suppose that we have training set $\mathcal{D}$ and testing set $\mathcal{T}$. Under closed-world assumption, both training and testing data are from in-distribution $\mathcal{I}$, *i.e.*, $\mathcal{D} = \mathcal{D}_L \subset \mathcal{I}$ and $\mathcal{T} = \mathcal{T}^I \subset \mathcal{I}$. The subscript L means that $\mathcal{D}_L$ is fully labeled. Samples in $\mathcal{T}$ ($\mathcal{T}^I$) only belong to known classes $\mathcal{C}^I$ that provided by labels in $\mathcal{D}$ ($\mathcal{D}_L$). However, a more realistic setting suggests that $\mathcal{T}$ also contains unknown classes $\mathcal{C}^O$ that from out-of-distribution $\mathcal{O}$, *i.e.*, $\mathcal{T}$ is composed by $\mathcal{T}^I \subset \mathcal{I}$ and $\mathcal{T}^O \subset \mathcal{O}$. A model trained by $\mathcal{D}$ is required to not only correctly classify samples from $\mathcal{T}^I$ into $\mathcal{C}^I$, but also recognize OOD samples from $\mathcal{T}^O$. To this end, an unlabeled set $\mathcal{D}_U$ is introduced to assist the training process, leading to $\mathcal{D} = \mathcal{D}_L \cup \mathcal{D}_U$. Ideally, unlabeled set should be purely from out-of-distribution, *i.e.* $\mathcal{D}_U \subset \mathcal{O}$. However, in the real practice, $\mathcal{D}_U$ is a mixture of both $\mathcal{D}_U^I \subset \mathcal{I}$ and $\mathcal{D}_U^O \subset \mathcal{O}$ with unknown separation. It is also mentioned that $\mathcal{D}_U^O$ does not necessarily cover $\mathcal{T}^O$. In summary, our goal is to train an image classifier from training set $\mathcal{D} = \mathcal{D}_L \cup \mathcal{D}_U$, so that the model has the capacity to reject $\mathcal{T}^O$, in addition to classify samples from $\mathcal{T}^I$ correctly.

3.2. Framework Overview

To empower the classifier with OOD detection ability using both ID set $\mathcal{D}_L$ and unlabeled set $\mathcal{D}_U$, our pipeline design is initiated by a classic OE architecture [13] that trains the network to classify ID samples correctly and forces high-entropy predictions on unlabeled samples, which is encapsulated in the classifier branch to be introduced in Section 3.3. Then, an unsupervised dual grouping (UDG) is proposed to group $\mathcal{D}_L$ and $\mathcal{D}_U$ altogether. Based on the grouping, ID samples from $\mathcal{D}_U$ can be filtered out and redirected to $\mathcal{D}_L$ to enhance the performance of classification branch, which is introduced in Section 3.4. With the grouping information produced by UDG, an auxiliary deep clustering branch is attached, aiming to discover the valuable but understudied knowledge in the unlabeled set. Finally, the entire training and testing procedure is summarized in Section 3.5. Figure 2 illustrates the proposed pipeline.

3.3. Main Task: Classification and Entropy Loss

We firstly focus on the ID classification ability of the proposed model. A classifier is built, containing a backbone encoder E with learnable parameter θ_E and a classification head F_C with learnable parameter θ_C. A standard cross-entropy loss is utilized to train the classifier using data-label pairs in $\mathcal{D}_L$, formulated by Equation 1.

$$\mathcal{L}_C^L = -\frac{1}{|\mathcal{D}_L|} \sum_{(x_i, y_i) \in \mathcal{D}_L} \log \Big(p_{y_i}(y|x_i, \theta_E, \theta_C) \Big) \quad (1)$$

Now, we use the unlabeled set to help the network gain OOD detection capabilities, following the classic architecture in outlier exposure [13]. Ideally, all samples from the unlabeled set are OOD, *i.e.* $\mathcal{D}_U \subset \mathcal{O}$. In this case, since they do not belong to any one of the known classes, the network is forced to produce a uniform posterior distribution over all known classes for unlabeled samples. Therefore, an entropy loss is introduced in Equation 2 to flatten the model prediction on unlabeled samples [13].

$$\mathcal{L}_C^U = -\frac{1}{|\mathcal{D}_U|} \frac{1}{|\mathcal{C}_I|} \sum_{x_i \in \mathcal{D}_U} \sum_{c \in \mathcal{C}_I} \log \Big(p_c(y|x_i, \theta_E, \theta_C) \Big) \quad (2)$$

However, in real practice, $\mathcal{D}_U$ might be mixed with ID samples, leading to the assumption of $\mathcal{D}_U \subset \mathcal{O}$ inaccurate. The problem awaits solving by the proposed unsupervised dual grouping in Section 3.4.

3.4. Unsupervised Dual Grouping (UDG)

In this section, we first introduce the basic operation of UDG and then focus on how UDG solves the mentioned

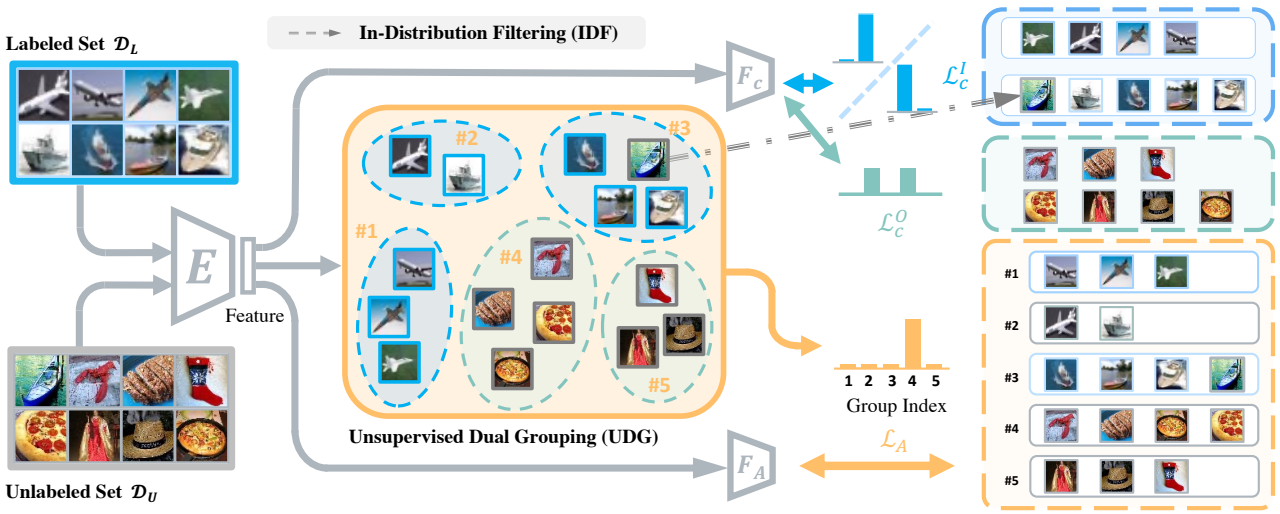


Figure 2: **The proposed framework of OOD detection with unsupervised dual grouping (UDG).** The CNN model has one encoder E and two fully-connected heads F_C and F_A. F_C is a classification head, which enables the model to correctly predict ID samples with classification loss $\mathcal{L}_C^I$, and to force flatten predictions on unlabeled samples using entropy loss $\mathcal{L}_C^O$. F_A is an auxiliary head for deep clustering. During the UDG process, the operation of in-distribution filtering (IDF) considers unlabeled samples that fall into ID groups as ID samples for later loss calculation. The group index is then used for the auxiliary deep clustering task.

problem on entropy loss $\mathcal{L}_C^U$ when the ID samples are mixed in $\mathcal{D}_U$. Ideally, samples in $\mathcal{D}_U^I$ should be removed from $\mathcal{D}_U$ and return to $\mathcal{D}_L$ with their corresponding labels, only leaving $\mathcal{D}_U^O$ for entropy loss minimization. Fortunately, UDG has the ability to complete the task with In-Distribution Filtering (IDF) operation. Besides, the grouping information provided by UDG can be used by an auxiliary branch of deep clustering to explore the knowledge in the unlabeled data and in turn boost the semantic capability of the model.

Basic Operation. We divide the entire training set $\mathcal{D}$ into K groups for every epoch. At the t-th epoch, the encoder E (denoted as $E^{(t)}$) extracts the feature of every training sample to form a feature set $\mathcal{F}^{(t)}$ for the following grouping process. Any clustering methods can be implemented for grouping, while in this work, we use the classic k-means algorithm [26] as in Equation 3, where $\mathcal{G}^{(t)}$ restores group index of every sample and $g_i^{(t)} \in \mathcal{G}^{(t)}$ denotes the group index of sample x_i at epoch t.

$$\mathcal{G}^{(t)} \leftarrow kmeans(\mathcal{F}^{(t)}), \text{ where } \mathcal{F}^{(t)} = \{E^{(t)}(x)|x \in \mathcal{D}\} \tag{3}$$

Since then, every sample should belong to one of the K groups. Formally, all samples that belong to k-th group at the t-th epoch form the set $\mathcal{D}_k$ in Equation 4.

$$\mathcal{D}_k^{(t)} = \{x_i|g_i^{(t)} = k, (x_i, g_i^{(t)}) \in (\mathcal{D}, \mathcal{G}^{(t)})\} \tag{4}$$

In-Distribution Filtering (IDF) with UDG. In this part, we introduce how we filter out ID samples on the scope of groups. Empirically, for a group that is dominated by a labeled class, the unlabeled data in the group is more likely to belong to the corresponding category. With the rule that the ID property of the sample can be estimated based on the group they belong to, the operation of In-Distribution Filtering (IDF) is proposed to filter out ID samples from the unlabeled set. For group k, we define its group purity $\gamma_{k,c}^{(t)}$ to show the proportion of samples belonging to class c at epoch t as Equation 5, where $[\mathcal{D}_L]_c$ denotes all labeled samples within class c.

$$\gamma_{k,c}^{(t)} = \frac{|\mathcal{D}_k^{(t)} \cap [\mathcal{D}_L]_c|}{|\mathcal{D}_k^{(t)}|} \tag{5}$$

With group purity $\gamma_{k,c}^{(t)}$ prepared, IDF operator returns all unlabeled samples in a group with group purity over a threshold τ back to labeled set with their labels identical to the group majority, forming an updated labeled set $\mathcal{D}_L^{(t)}$ according to Equation 6.

$$\mathcal{D}_L^{(t)} = \mathcal{D}_L \cup \{x|x \in \mathcal{D}_k^{(t)}, \gamma_{k,c}^{(t)} > \tau\} \tag{6}$$

Notice that from the second epoch, $\mathcal{D}_L$ in Equation 5 will be replaced with $\mathcal{D}_L^{(t-1)}$, but $\mathcal{D}_L$ in Equation 6 remains unchanged to enable an error correction mechanism.

Complementary to the labeled set which is updated into $\mathcal{D}_L^{(t)}$, unlabeled set is also updated as $\mathcal{D}_U^{(t)}$. Using the updated sets, both the classification loss and entropy loss are

modified as Equation 7 and Equation 8.

$$[\mathcal{L}_C^I]^{(t)} = -\frac{1}{|\mathcal{D}_L^{(t)}|} \sum_{(x_i,y_i)\in\mathcal{D}_L^{(t)}} \log\Big(p_{y_i}(y|x_i,\theta_E,\theta_C)\Big) \tag{7}$$

$$[\mathcal{L}_C^O]^{(t)} = -\frac{1}{|\mathcal{D}_U^{(t)}|}\frac{1}{|\mathcal{C}_I|} \sum_{x_i\in\mathcal{D}_U^{(t)}} \sum_{c\in\mathcal{C}_I} \log\Big(p_c(y|x_i,\theta_E,\theta_C)\Big) \tag{8}$$

Auxiliary Task for UDG. The motivation for designing the auxiliary branch is to fully leverage the knowledge contained in the unlabeled set, expecting that the learned semantics can further benefit the model performance, especially on ID classification. It is expected that the learned semantics can further benefit the model performance, especially on ID classification. Fortunately, the groups that UDG provides are perfectly compatible with a deep clustering process [18], which is used as the unsupervised auxiliary task for knowledge exploration.

The intuition of deep clustering is that samples that lie in the same group are supposed to be in the same category. As the group index for every sample is provided in $\mathcal{G}^{(t)}$ by Equation 3, a fully-connected auxiliary head with learnable parameter θ_A is trained to classify the sample into its corresponding group with auxiliary loss $\mathcal{L}_A$ in Equation 9.

$$\mathcal{L}_A^{(t)} = -\frac{1}{|\mathcal{D}|} \sum_{(x_i,g_i)\in(\mathcal{D},\mathcal{G}^{(t)})} \log\Big(p_{g_i}(y|x_i,\theta_E,\theta_A)\Big) \tag{9}$$

3.5. Training and Testing Process

Finally, with the modified classification loss $[\mathcal{L}_C^I]^{(t)}$, modified entropy loss $[\mathcal{L}_C^O]^{(t)}$ and auxiliary loss $\mathcal{L}_A^{(t)}$, the final loss $\mathcal{L}$ can be calculated by Equation 10 with hyper-parameter λ_U and λ_A. An end-to-end training process is performed to optimize encoder E with parameter θ_E, classification head F_C with parameter θ_C and auxiliary head F_A with parameter θ_A simultaneously.

$$\mathcal{L}^{(t)} = [\mathcal{L}_C^I]^{(t)} + \lambda_U \cdot [\mathcal{L}_C^O]^{(t)} + \lambda_A \cdot \mathcal{L}_A^{(t)} \tag{10}$$

During testing, only classification head F_C along with backbone encoder E is utilized. The model will only make an in-distribution prediction if the value of maximum prediction passes a pre-defined threshold δ. Otherwise, the sample would be considered as out-of-distribution one. The testing process is formalized by Equation 11.

$$pred = \begin{cases} \textit{N.A.}, & \text{if } \max p(y|x_i,\theta_E,\theta_C) < \delta, \\ \arg\max_c p_c(y|x_i,\theta_E,\theta_C), & \text{otherwise.} \end{cases} \tag{11}$$

Table 2: **Tiny-ImageNet classes that are semantically coherent with exemplar CIFAR-10 classes.** All images in these classes are labeled as ID for SC-OOD benchmarking.

CIFAR-10	Tiny-ImageNet Classes	
cat	n02802426	Tabby, Tabby Cat
	n02977058	Egyptian Cat
	n04146614	Persian Cat
dog	n02823428	Golden Retriever
	n03388043	Chihuahua
	n02056570	Yorkshire Terrier
	n03891332	Labrador Tetriever
	n03042490	German Shepherd
	n03930313	Standard Poodle

Figure 3: **Exemplar ID images hidden in irrelevant categories are also filtered for SC-OOD CIFAR-10 benchmark.** Although the class-wise filtering by Table 2 can identify a large portion of ID samples, some multi-label images that contains ID semantics requires manually filtering.

4. SC-OOD Benchmarks

In this section, we introduce two benchmarks to reflect semantically coherent OOD detection. Two benchmarks consider two famous datasets of CIFAR-10/100 [27] as in-distribution, respectively. Five other datasets including Texture [28], SVHN [29], Tiny-ImageNet [30], LSUN [31], and Places365 [32] are prepared as OOD datasets. We re-split $\mathcal{T}^I$ and $\mathcal{T}^O$ according to the semantics of samples for SC-OOD benchmarks. There are two steps for re-splitting: **1)** we first pick out the ID classes from the OOD datasets and mark all the images within the selected classes as ID samples. Table 2 shows exemplar Tiny-ImageNet ID classes corresponding to two CIFAR-10 classes. **2)** we then conduct fine-grained filtering since many images from irrelevant OOD categories also contain ID semantics. Figure 3 shows exemplar Tiny-ImageNet ID images that are hidden by irrelevant labels. Eventually, we obtain CIFAR-10/100 SC-OOD benchmarks with detailed descriptions as follows.

4.1. Benchmark for CIFAR-10

CIFAR-10 is a natural object image dataset with 50,000 training samples and 10,000 testing samples from 10 object classes. Selected datasets for $\mathcal{T}$ include **1)** CIFAR-10

test set with all 10,000 images as $\mathcal{T}^I$; **2)** Entire Texture set with 5,640 images of textural images, all as $\mathcal{T}^O$; **3)** SVHN test set with 26,032 images of real-world street numbers, all as $\mathcal{T}^O$; **4)** CIFAR-100 test set with 10,000 object images that disjoint from CIFAR-10 classes, therefore all as $\mathcal{T}^O$; **5)** Tiny-ImageNet test set containing 10,000 images from 200 objects, with 1,207 images as $\mathcal{T}^I$ and 8,793 images as $\mathcal{T}^O$; **6)** LSUN test set containing 10,000 images for scene recognition, with 2 images as $\mathcal{T}^I$ and 9,998 images as $\mathcal{T}^O$; **7)** Places365 test set containing 36,500 scene images, with 1,305 images as $\mathcal{T}^I$ and 35,195 images as $\mathcal{T}^O$.

4.2. Benchmark for CIFAR-100

CIFAR-100 is a dataset of 100 fine-grained classes with 50,000 training samples and 10,000 testing samples. The classes between CIFAR-10 and CIFAR-100 are disjoint. Selected datasets for $\mathcal{T}$ include **1)** CIFAR-100 test set with all 10,000 images as $\mathcal{T}^I$; **2)** Entire Texture set with 5,640 images of textural images, all as $\mathcal{T}^O$; **3)** SVHN test set with 26,032 images of real-world street numbers, all as $\mathcal{T}^O$; **4)** CIFAR-10 test set with 10,000 object images that disjoint from CIFAR-100 classes, therefore all as $\mathcal{T}^O$; **5)** Tiny-ImageNet test set with 2,502 images as $\mathcal{T}^I$ and 7,498 images as $\mathcal{T}^O$; **6)** LSUN test set with 2,429 images as $\mathcal{T}^I$ and 7,571 images as $\mathcal{T}^O$; **7)** Places365 with 2,727 images as $\mathcal{T}^I$ and 33,773 images as $\mathcal{T}^O$.

4.3. Evaluation Metrics

We use four kinds of metrics to evaluate the performance on both ID classification and OOD detection.

FPR95 is short for the false positive rate (FPR) at 95% true positive rate (TPR). It measures the portion of falsely recognized OOD when the most of ID samples are recalled.

AUROC computes the area under the receiver operating characteristic curve, evaluating the OOD detection performance. Samples from $\mathcal{T}^I$ are considered positive.

AUPR measures the area under the precision-recall curve. Depending on the selection of positiveness, AUPR contains AUPR-In, which regards $\mathcal{T}^I$ as positive, as well as AUPR-Out, where $\mathcal{T}^O$ is considered as positive. In Table 1 and Table 3, we use AUPR to represent the value of AUPR-Out due to its complementary polarities with AUROC.

CCR@FPRn shows Correct Classification Rate (CCR) at the point when FPR reaches a value n. The metric evaluates ID classification and OOD detection simultaneously and is formalized by Equation 3 in [12].

Among all the mentioned metrics, only FPR95 is expected to have a lower value on a better model. Higher values on any other metrics indicate better performance.

5. Experiments

In this section, after describing the implementation details, the effect of each component is analyzed in the ablation study. Then we compare our method with previous state-of-the-art methods. Finally, a more in-depth exploration of our method is discussed.

Experimental Settings. Two experimental sets are performed with $\mathcal{D}_L$ of CIFAR-10 and CIFAR-100 [27], respectively. Both of the training set contains 50,000 images. Tiny-ImageNet [30] training set is used as $\mathcal{D}_U$ in both experiments. Testing is conducted according to Section 4. The main paper only reports the average metric values on all 6 datasets in each benchmark. All ablation and analytical experiments are performed on the CIFAR-10 benchmark.

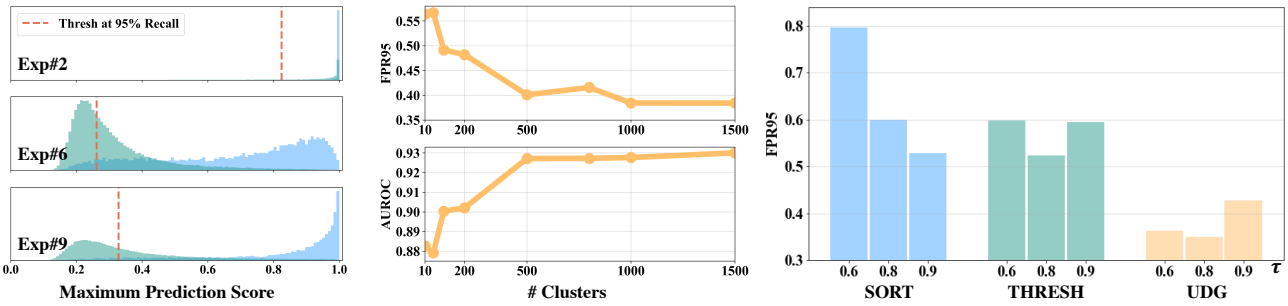
Implementation Details. All experiments are performed with a standard ResNet-18 [33], trained by an SGD optimizer with a weight decay of 0.0005 and a momentum of 0.9. Two data-loaders are prepared with batch size of 128 for $\mathcal{D}_L$ and 256 for $\mathcal{D}_U$. A cosine learning rate scheduler is used with an initial learning rate of 0.1, taking totally 100 epochs. For hyperparameters of UDG, we set $\lambda_U = 0.5$ and $\lambda_A = 0.1$ for all experiments. Group number K for CIFAR-10/100 is 1000/2000 with IDF threshold $\tau = 0.8$.

5.1. Ablation Study

In this section, we will analyze the effect of every major component, including classification task $\mathcal{L}_C^{\mathcal{I}}$, OE loss $\mathcal{L}_C^{\mathcal{U}}$, auxiliary deep clustering task $\mathcal{L}_{\mathcal{A}}$, and in-distribution filtering (IDF) operator F by Table 3. In summary, OE loss $\mathcal{L}_C^{\mathcal{U}}$ and IDF operator are shown most effective for OOD detection, while the auxiliary deep clustering task can further improve the performance. Notably, we also report the basic classification accuracy on CIFAR-10 test set, denoted as ACC. Detailed explanation is conveyed as follows.

Effectiveness of Unlabeled Data. Table 3 is divided into two major blocks according to the usage of unlabeled Tiny-ImageNet. In this part, we discuss the differences brought by the introduction of unlabeled data. Exp#2 operates the standard classification where only $\mathcal{L}_C^I$ is used, and Exp#6 is the standard OE method [13] using an additional $\mathcal{L}_U^O$. The result shows that the OOD detection ability is enhanced by a large margin, as FPR95 gets a significant 7.74% improvement. Figure 4-a compares the histogram of maximum prediction scores between Exp#2 and #6. The overconfident property is largely reduced in favor of unlabeled samples, and the ID/OOD discrepancy is also enlarged. However, it is also worth noting that the ID classification accuracy is reduced from 94.94% to 91.87%. Since we use a realistic unlabeled set mixing both ID and OOD for OE loss, the unlabeled ID data will wrongly contribute to $\mathcal{L}_C^{\mathcal{O}}$ and therefore harm the classification performance.

Analysis of Unsupervised Dual Grouping. The contribution of UDG is twofold: **1)** enabling the IDF operation; **2)** creating an auxiliary loss $\mathcal{L}_A$. In this part, we especially focus on the analysis of $\mathcal{L}_A$. Exp#1 operates a standard deep clustering on CIFAR-10 ($K = 50$). The fully-connected



(a) Histogram of Maximum Prediction Score (b) Influence of Cluster Numbers (c) Comparison between different IDF Designs

Figure 4: **Comparison and analysis to demonstrate the effectiveness of each module in our framework.** (a) shows the discrepancy between the maximum prediction scores of ID (blue) and OOD (green) samples in three experiments of Table 3 statistically. (b) shows the the proposed method can obtain a stably good result with a large number of pre-defined clusters. (c) shows the in-distribution filtering (IDF) strategy we use, *i.e*. UDG, is significantly better than the alternatives.

layer is finetuned on training set at the end of the training. A distressing result is obtained on all metrics. Even worse is Exp#4, which shows that expanding the training set with unlabeled data dominated by OOD further destroys the OOD detection ability of fully unsupervised methods. Possible explanation is that without the knowing the ID/OOD, deep clustering may easily group ID/OOD samples into one cluster, thus giving up the OOD detection capability. Therefore, although Exp#2 and #3 show that on the standard CIFAR training set, the auxiliary $\mathcal{L}_A$ can benefit all metrics, the comparison between Exp#3 and Exp#5 unfortunately shows that once an OOD-mixed unlabeled data involved, a simple combination of classification task and unsupervised deep clustering task will damage the OOD detection ability. Fortunately, after introducing the OOD discrepancy loss $\mathcal{L}_U^O$, the comparison between Exp#6 and Exp#7 shows that the disadvantage of $\mathcal{L}_A$ can be largely reduced, rekindling our hope of regaining the value of $\mathcal{L}_A$.

Effectiveness of In-Distribution Filtering. The contribution of IDF is in twofold: **1)** improving ID classification by collecting ID samples from the unlabeled set for better $\mathcal{L}_C^I$, and **2)** purify from unlabeled ID/OOD mixture into a clean OOD set for better $\mathcal{L}_U^O$. The comparison between Exp#6 and #8 (UDG but $\lambda_A = 0$) illustrates that IDF completes the above goals with a significant 10.4% benefit on FPR95 and an improvement on classification accuracy. By using a cleaner ID and OOD sets, the auxiliary $\mathcal{L}_A$ finally becomes beneficial in the completed version of Exp#9.

5.2. Benchmarking Results

Table 4 compares our proposed approach with previous state-of-the-art OOD detection methods. Here we only report the average metric values on all 6 OOD datasets for each benchmark due to limited space. Full results are shown in Appendix. Results show that our proposed UDG achieves better results on both SC-OOD benchmarks.

Table 3: **Ablation study on SC-OOD CIFAR-10 benchmark** to show the effectiveness of every component in the proposed framework. F represents for the IDF operator. For simplicity, we refer to each experiment based on its index (*e.g*. Exp#2 for the experiment with $\mathcal{L}_C^{\mathcal{I}}$ only).

$\mathcal{D}$	Components	FPR95	AUROC	AUPR	ACC
CIFAR	1: $\mathcal{L}_{\mathcal{A}}$	89.53	65.80	65.46	64.88
	2: $\mathcal{L}_C^{\mathcal{I}}$	58.27	89.25	87.72	94.94
	3: $\mathcal{L}_C^{\mathcal{I}} + \mathcal{L}_{\mathcal{A}}$	**55.62**	**90.72**	**88.33**	**95.02**
CIFAR+TIN	4: $\mathcal{L}_{\mathcal{A}}$	91.15	64.00	63.47	59.02
	5: $\mathcal{L}_C^{\mathcal{I}} + \mathcal{L}_{\mathcal{A}}$	**62.75**	**88.21**	**86.45**	**94.68**
	6: $\mathcal{L}_C^{\mathcal{I}} + \mathcal{L}_C^{\mathcal{U}}$	50.53	88.93	87.83	91.87
	7: $\mathcal{L}_C^{\mathcal{I}} + \mathcal{L}_C^{\mathcal{U}} + \mathcal{L}_{\mathcal{A}}$	51.41	90.53	88.17	90.70
	8: $\mathcal{L}_C^{\mathcal{I}} + \mathcal{L}_C^{\mathcal{U}} + F$	40.93	92.23	91.92	92.34
	9: $\mathcal{L}_C^{\mathcal{I}} + \mathcal{L}_C^{\mathcal{U}} + F + \mathcal{L}_{\mathcal{A}}$	**36.22**	**93.78**	**92.61**	**92.94**

ODIN [7] and **Energy-based OOD detector (EBO)** [15] are two representative post-processing OOD methods. We report their best results after hyperparameter searching. Their performances are usually inferior to OE method.

Outlier Exposure (OE) [13] corresponds to Exp#6 in Section 3. The results show that using OOD in training with this mechanism can gain an advantage beyond other baselines.

Maximum Classifier Discrepancy (MCD) enlarges the entropy discrepancy between two branches to detect OODs [14]. However, we find it significantly overfits the training OOD samples while is difficult to generalize to other OOD domains, leading to disappointing results.

UDG achieves state-of-the-art results on all metrics of OOD detection. In particular, FPR@95 is significantly reduced on both benchmarks. The full table in the appendix shows that UDG can not only have an advantage on revised SC-OOD datasets such as CIFAR-Places365, but also benefits classic

Table 4: **Comparison between previous state-of-the-art methods and ours on the SC-OOD CIFAR-10/100 benchmarks.** All experiments use ResNet-18 [33] for fair comparison. ODIN [7] and EBO [15] do not require external data, and OE [13], MCD [14], and our UDG use Tiny-ImageNet as unlabeled data. UDG obtains consistently better results on almost all metrics.

$\mathcal{D}_I$ ($\mathcal{D}_U$)	Method	FPR95 ↓	AUROC ↑	AUPR(In/Out) ↑	CCR@FPR ↑			
					10^{-4}	10^{-3}	10^{-2}	10^{-1}
CIFAR-10 (Tiny-ImageNet)	ODIN [7]	52.00	82.00	73.13 / 85.12	0.36	1.29	6.92	39.37
	EBO [15]	50.03	83.83	77.15 / 85.11	0.49	1.93	9.12	46.48
	OE [13]	50.53	88.93	87.55 / 87.83	13.41	20.25	33.91	68.20
	MCD [14]	73.02	83.89	83.39 / 80.53	5.41	12.3	28.02	62.02
	UDG (ours)	**36.22**	**93.78**	**93.61 / 92.61**	**13.87**	**34.48**	**59.97**	**82.14**
CIFAR-100 (Tiny-ImageNet)	ODIN [7]	81.89	77.98	78.54 / 72.56	1.84	5.65	17.77	46.73
	EBO [15]	81.66	79.31	80.54 / 72.82	2.43	7.26	21.41	**49.39**
	OE [13]	80.06	78.46	80.22 / 71.83	2.74	8.37	**22.18**	46.75
	MCD [14]	85.14	74.82	75.93 / 69.14	1.06	4.60	16.73	41.83
	UDG (ours)	**75.45**	**79.63**	**80.69 / 74.10**	**3.85**	**8.66**	20.57	44.47

OOD detection test dataset pairs such as CIFAR-Texture, which does not have semantic conflicts.

5.3. Further Analysis

Influence of Cluster Numbers. Figure 4-b shows the influence of the pre-defined cluster number K. Generally, increasing K helps converge to the optimal result. When K is small, a larger group size will prevent any group from completely belonging to one class, making it difficult for IDF to filter out any ID samples. Also, large groups will inevitably include both ID and OOD samples, leading the deep clustering task to obscure the ID/OOD discrepancy. Therefore, our proposed method requires a certain large number of clusters. Fortunately, experiments show that the results are insensitive to the number of clusters when it is large ($K \geq 500$), reflecting the practicality of our method.

The Design Choice of IDF. Apart from the IDF proposed in Section 3.4, there are two straightforward alternatives for ID sample filtering. One solution (denoted as "THRESH") filters all unlabeled samples whose maximum softmax scores exceed a predefined threshold τ. Another solution (called "SORT") sorts all unlabeled samples based on their maximum softmax scores and selects the top $(1-\tau)\%$ samples as new ID samples. We also denote our proposed group-based IDF strategy as "UDG", which takes all unlabeled samples in a group with ID purity over τ as ID samples. Figure 4-c shows the comparison between them. Generally, our proposed UDG obtains the best performance on FPR95 comparing to other IDF strategies. SORT achieves the worst performance, since it directly includes a certain number of unlabeled images as ID from the beginning of the end-to-end training. The wrongly introduced samples will join classification task so that the error could accumulate, preventing the model from adopting the OOD detection ability properly. Smaller τ will let SORT include more unlabeled images without the guarantee of the filtering accuracy. THRESH has better performance due to its better control of unlabeled samples inclusion. However, it is still not comparable to UDG, which takes ID samples in a more conservative manner by exploiting a grouping mechanism to reduce overconfidence characteristics of the deep models. The results indicate that the group-based ID filtering performs more stably than sample-based methods.

6. Conclusion

In this paper, we highlight a problem of current OOD benchmarks which split ID/OOD according to the data source rather than the semantic meaning, and therefore redesign realistic and challenging SC-OOD benchmarks. An elegant pipeline named UDG is proposed to achieve the state-of-the-art result on SC-OOD benchmarks, with the usage of realistic unlabeled set. We hope that the more realistic and challenging SC-OOD setting provides new research opportunities for the OOD community and draw researchers' attention to the importance of data review.

Acknowledgments

This work was supported by Innovation and Technology Commission of the Hong Kong Special Administrative Region, China (Enterprise Support Scheme under the Innovation and Technology Fund B/E030/18), NTU NAP, and RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015. 1

[2] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *IJCV*, 2015. 1

[3] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *arXiv preprint arXiv:2103.02503*, 2021. 1

[4] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *ICLR*, 2017. 1, 2

[5] X. Huang, D. Kroening, W. Ruan, J. Sharp, Y. Sun, E. Thamo, M. Wu, and X. Yi, "A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability," *Computer Science Review*, 2020. 1

[6] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The kitti vision benchmark suite," in *CVPR*, 2012. 1

[7] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," in *ICLR*, 2017. 1, 2, 7, 8, 10

[8] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *NeurIPS*, 2018. 1, 2

[9] J. Serrà, D. Álvarez, V. Gómez, O. Slizovskaia, J. F. Núñez, and J. Luque, "Input complexity and out-of-distribution detection with likelihood-based generative models," in *ICLR*, 2020. 1, 2

[10] H. Choi, E. Jang, and A. A. Alemi, "WAIC, but why? Generative ensembles for robust anomaly detection," *arXiv preprint arXiv:1810.01392*, 2018. 1, 2

[11] A. Vyas, N. Jammalamadaka, X. Zhu, D. Das, B. Kaul, and T. L. Willke, "Out-of-distribution detection using an ensemble of self supervised leave-out classifiers," in *ECCV*, 2018. 1, 2

[12] A. R. Dhamija, M. Günther, and T. Boult, "Reducing network agnostophobia," in *NeurIPS*, 2018. 2, 6

[13] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," in *ICLR*, 2019. 2, 3, 6, 7, 8

[14] Q. Yu and K. Aizawa, "Unsupervised out-of-distribution detection by maximum classifier discrepancy," in *ICCV*, 2019. 2, 3, 7, 8

[15] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," in *NeurIPS*, 2020. 2, 7, 8

[16] J. Ren, P. J. Liu, E. Fertig, J. Snoek, R. Poplin, M. Depristo, J. Dillon, and B. Lakshminarayanan, "Likelihood ratios for out-of-distribution detection," in *NeurIPS*, 2019. 2

[17] Y. Li and N. Vasconcelos, "Background data resampling for outlier-aware classification," in *CVPR*, 2020. 3

[18] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *ECCV*, 2018. 3, 5

[19] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *CVPR*, 2015. 3

[20] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *ECCV*, 2016. 3

[21] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *ECCV*, 2016. 3

[22] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *ICLR*, 2018. 3

[23] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018. 3

[24] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020. 3

[25] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *ICML*, 2020. 3

[26] S. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, 1982. 4

[27] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," *Citeseer*, 2009. 5, 6

[28] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *CVPR*, 2014. 5

[29] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *Proceedings of NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011. 5

[30] Y. Le and X. Yang, "Tiny imagenet visual recognition challenge." `http://cs231n.stanford.edu/tiny-imagenet-200.zip`, 2015. 5, 6

[31] F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao, "LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv preprint arXiv:1506.03365*, 2015. 5

[32] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *TPAMI*, 2017. 5

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016. 6, 8, 11