LUNA: Localizing Unfamiliarity Near Acquaintance for Open-Set Long-Tailed Recognition

Jiarui Cai[1], Yizhou Wang[1], Hung-Min Hsu[1], Jenq-Neng Hwang[1], Kelsey Magrane[2], Craig Rose[3]

[1] University of Washington
[2] Pacific States Marine Fisheries Commission (PSMFC)
[3] FishNext Research
{jrcai, ywang26, hmhsu, hwang}@uw.edu, kelsey.magrane@noaa.gov, fishnextresearch@gmail.com

Abstract

The predefined artificially-balanced training classes in object recognition have limited capability in modeling real-world scenarios where objects are imbalanced-distributed with unknown classes. In this paper, we discuss a promising solution to the Open-set Long-Tailed Recognition (OLTR) task utilizing metric learning. Firstly, we propose a distribution-sensitive loss, which weighs more on the tail classes to decrease the intra-class distance in the feature space. Building upon these concentrated feature clusters, a local-density-based metric is introduced, called Localizing Unfamiliarity Near Acquaintance (LUNA), to measure the novelty of a testing sample. LUNA is flexible with different cluster sizes and is reliable on the cluster boundary by considering neighbors of different properties. Moreover, contrary to most of the existing works that alleviate the open-set detection as a simple binary decision, LUNA is a quantitative measurement with interpretable meanings. Our proposed method exceeds the state-of-the-art algorithm by 4-6% in the closed-set recognition accuracy and 4% in F-measure under the open-set on the public benchmark datasets, including our own newly introduced fine-grained OLTR dataset about marine species (MS-LT), which is the first naturally-distributed OLTR dataset revealing the genuine genetic relationships of the classes.

Introduction

There is a wide range of real-life object recognition tasks that operates under the training-learning paradigm and can be naturally modeled as the image classification task. Technical revolution sweeps various fields like species identification (Van Horn et al. 2018), medical imaging perception (Wang et al. 2020a), human face recognition (Deng et al. 2019) and scene classification in autonomous driving (Narayanan, Dwivedi, and Dariush 2019). However, the performances of the state-of-the-art object recognition methods mostly bias on the sample-rich classes that have been seen in the training set, with a limited ability on classifying the sample-few classes, not to mention the new/novel classes of objects (Kang et al. 2019; Zhou et al. 2020).

The main culprit of this phenomenon is the simulation scenario in the laboratory cannot fully model reality: the conformity between training and testing sets determines the system's dynamic performance, reliability and scalability. Back in the real world, the factual object samples are also unevenly distributed, and the object classes are always open-ended. The state-of-the-art algorithms are either focused on solving open-set issues (Bendale and Boult 2016; Scheirer, Jain, and Boult 2014; Ge et al. 2017), or only aimed to classify the objects under a closed long-tail distribution (Kang et al. 2019; Zhou et al. 2020). However, when there is a call for implementing object recognition in daily life, the open-set and long-tail challenges commonly coincide (Van Horn et al. 2018; Wellmer and Becker-Platen 2000; dis 2019). Separating them is twice the effort for half the result. To step closer to reality, Liu et al. attempt to merge and deal with the open-set long-tail recognition (OLTR) together by one framework in 2019 and proposed the OLTR baseline (Liu et al. 2019). However, existing OLTR approaches (Liu et al. 2019; Zhu and Yang 2020) still see some fundamental and methodological gaps:

- **No authentic collected open long-tail dataset to evaluate the OLTR methodology** (Liu et al. 2019): existing benchmarks are limited to artificially-sampled ones. The generic relationship among objects are disrupted. For example, there are only 9 samples for truck while 516 samples for white shark in ImageNet-LT.
- **Decoupling open set challenges with long-tailed distribution** (Bendale and Boult 2016; Ge et al. 2017; Júnior et al. 2017; Hassen and Chan 2020): when people study the open set issues, models are designed based on balanced sets (Deng et al. 2009; Zhou et al. 2017), which reduce the utility and value of transferring the methodology to OLTR tasks.
- **Exhaustively engage into the long-tail recognition and ignore the open set issues** (Cao et al. 2019; Huang et al. 2016; Lin et al. 2017; Cui et al. 2019; Li et al. 2020): due to the hurdle of recognition of objects from the imbalanced set, the literature focus on improving the accuracy in the long-tailed distribution without considering the open set scenario, which needs accommodations for the actual OLTR tasks.

The above three significant hurdles motivate us to research in the metric domain, which automates the feature selection and learns task-specific distance functions to access similarity (Kulis et al. 2012). In this paper, we propose a metric learning framework, called Localizing Unfamiliarity Near Acquaintance (LUNA), to quantitatively measure the level of novelty based on the local density of the deep CNN features for the open-set long-tailed recognition task. With LUNA, two questions can be answered precisely, (1) whether the input is novel or not; (2) if no, which class it is; if yes, what is the unfamiliarity level of the new class concerning the pretrained acquaintance classes. In summary, we claim our contributions and technical innovations as follows,

- We collect a new well-annotated real Marine Species open long-tailed (MS-LT) dataset. As the first natural OLTR dataset in a fine-grained domain, it will be a solid supplement to the existing manually re-sampled OLTR datasets. It poses new challenges on representation learning and novel species detection.
- To make the categories concentrated in feature space individually, the feature extractor is trained by a newly proposed loss, called weighted center loss, to minimize their intra-class distances so as to form dense clusters in the high-dimensional space. It centralizes the deep features of the head classes, while preserving the classification accuracy of the tails, resulting in more distinctive features.
- To measure the unfamiliarity level of the new classes and evaluate the closeness with acquaintance classes, we propose a LUNA factor, an outlier metric based on the relative density of the deep features, which is adaptive to the distribution. The LUNA factor is the first indicator that provides quantitative measurements of novelty under the long-tailed distribution.
- We extensively evaluate LUNA on the MS-LT dataset and two commonly-used artificially-sampled datasets, ImageNet-LT and Place-LT, in both long-tailed and open-set recognition tacks. The result shows that the LUNA significantly outperforms the state-of-the-art methods by 4-6% on the closed set and in average 4% improvement of the F-measure under the open-set setting.

Related Works

Long-tailed and open-set recognition. For long-tailed recognition, strategies have been explored to eliminate the bias towards heads in several ways, including one-stage re-balancing and multi-stage retraining. The re-balancing techniques consists of data re-sampling (either down-sampling of head classes (Galar et al. 2013; Liu, Wu, and Zhou 2008) or over-sampling of tail classes (Chawla et al. 2002; Han, Wang, and Mao 2005; Nguyen, Cooper, and Kamei 2011)) and loss re-weighting (Cao et al. 2019; Huang et al. 2016; Lin et al. 2017; Li et al. 2020). Though improving the performance on tails, however, such techniques hurt the model's generalizability and overall feature learning, leaving the heads under-represented. Multi-stage methods (Cui et al. 2019; Kang et al. 2019; Wang et al. 2020b) overcome it by decoupling the biased representation learning and re-balanced classifier training, achieving state-of-the-art performance. Nonetheless, they confine only within the closed set, without considering the adaption to unseen categories.

On the other hand, as it is impossible to collect all novel categories other than the trained ones, open-set recognition is of significant practical value. There are mainly three categories: (1) separate novel sampling with Weibull fitting (Scheirer, Jain, and Boult 2014; Bendale and Boult 2016); (2) train with synthesized samples (Ge et al. 2017); (3) cluster known categories in the feature space (Júnior et al. 2017; Hassen and Chan 2020). The idea of our proposed weighted center loss is consistent with the last type. The existing algorithms are evaluated on relatively small and balanced datasets, like MNIST (mni 1998). Besides, they utilize the naive distance of an instance to its closest class center as the outlier score, and the out-of-distribution known samples are prone to be classified as novelties.

Liu et al. (Liu et al. 2019) formally define the OLTR problem and set up benchmarks for evaluation. They also develop an OLTR network that dynamically learns meta-embedding for training samples as a combination of direct features from CNN and memory features representing the class-specific feature centroids to transfer knowledge from head to tail classes. The minimum confidence from a cosine classifier indicates the novelty of testing samples. However, the OLTR network uses distance to model reachability, which might be ambiguous where the clusters are overlapped. Also, end-to-end training decreases its interpretability.

Novelty Detection. Novelty detection, or anomaly detection, is a binary classification problem that aims to detect the outliers given few or no annotations on novel classes. Besides the above open-set related methods, the local outlier factor (LOF) is widely used (Breunig et al. 2000; Kriegel et al. 2009; Wang and Lu 2011; Zhu et al. 2018). In general, it compares the density of a certain data point with its neighboring K points (Breunig et al. 2000; Tu et al. 2018). Since the LOF is not flexible with different cluster sizes, in the long-tailed scenario, small K may limit the accuracy on head classes that occupy a larger portion of feature space, while local outliers may be ignored if a large K is chosen. The proposed LUNA is size-sensitive and it improves the reliability of the cluster boundaries by comparing with different groups of local neighbors.

Deep Metric Learning (DML). DML is to maximize the inter-class distances and minimize the intra-class distances in the high-dimensional embedding space. Two types of DML methods are widely used: a) learning with class-level labels and b) image-level labels. The former obtains embeddings from a classification model, e.g., ArcFace (Deng et al. 2019), CosFace (Wang et al. 2018). The latter optimizes the embedding distance of sampled image pairs or groups by loss functions directly, without generating the classification outcomes after DML, such as contrastive (Chopra, Hadsell, and LeCun 2005), triplet (Schroff, Kalenichenko, and Philbin 2015) and center (Wen et al. 2016) loss. These DML algorithms are all within the hypothesis that training data is sufficient and generally balanced, which does not hold for the long-tailed setting. For class-level DML, the classification accuracy is mostly affected by biases in data distribution; while for image-level DML, few-shot classes are easily over-fitted. In this paper, a frequency-aware loss function is proposed to tackle the data imbalance and metric learning problem simultaneously.

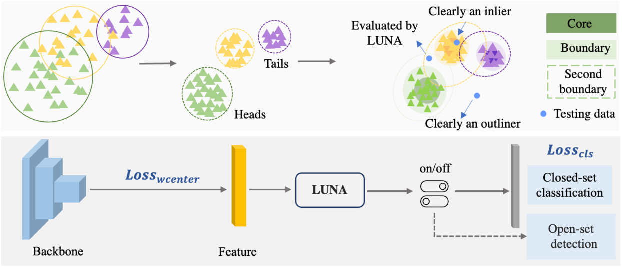

Figure 1: The illustration of our proposed method. Bottom: the workflow of open long-tailed recognition training and inference. Top: a brief illustration of our open-set detection method in feature space. The training samples form clusters in the feature space with the wcenter loss are further categorized into core, boundary and second boundary points by their relative local density. LUNA assesses the aforementioned metrics to measure the novelty of the testing samples.

The Proposed OLTR Method

In this section, we introduce our approach on training the classification network with a proposed distribution-sensitive loss to obtain distinctive representations, and detect novel classes based on the Localizing Unfamiliarity Near Acquaintance (LUNA, detailed illustration in Figure 1) measurement under the open-set setting.

Representation Learning: WCenter Loss

The center loss (Wen et al. 2016) is originally proposed for face recognition, which is formulated as

$$L_c = \frac{1}{2}\sum_i ||\mathbf{x}_i - \mathbf{c}_{y_i}||^2, \tag{1}$$

where $\mathbf{x}_i$ indicates the feature of the i-th sample with ground truth label y_i; $\mathbf{c}_{y_i}$ is the corresponding centroid, which is initialized randomly and updated iteratively to minimize the distance between itself and the continuously updated deep features during the training. Center loss is jointly trained with cross-entropy loss, balanced by a scalar parameter λ

$$L = L_{xent} + \lambda L_c, \tag{2}$$

where different λ ($\lambda = 0.001, 0.01, 0.1, 1$) is shown to lead to different deep feature distributions (Wen et al. 2016), and features are more concentrated with larger λ. In the situation of long-tailed datasets, the tail classes tend to be sparser in distribution since there are much fewer samples, which are easier to mix up with other clusters in the feature space. Thus, we propose a weighted center (wcenter) loss that caters to imbalanced distribution.

$$\begin{aligned} L_{wc} &= \frac{1}{2}\sum_i \lambda_i ||\mathbf{x}_i - \mathbf{c}_{y_i}||^2 \\ &= \frac{1}{2}\sum_i \left(\frac{\widetilde{n_j}}{\max_c\{\widetilde{n_c}\}} + 1\right) \cdot ||\mathbf{x}_i - \mathbf{c}_{y_i}||^2, \end{aligned} \tag{3}$$

where $\widetilde{n_j} = \frac{\max_c\{n_c\}}{n_j}$. Here, c is the category index; λ_i is the weight of normalized frequency of the class j that y_i belongs to; n_j denotes the number of samples of class j in the training set. Basically, λ_i is the inversed distribution scaled by the maximum frequency value, which is denoted as $\widetilde{n_j}$ and then normalized between $[0, 1]$. The fewer samples in a class, the higher weight it gets. Note that λ_i should be greater than 1 and at the same scale of cross-entropy loss to ensure convergence of the network. Therefore, we empirically add 1 to adjust the scale. Experiments in (Wen et al. 2016) also support this conclusion. Overall, the objective function is

$$L = L_{xent} + L_{wc}. \tag{4}$$

Note that the parameter λ is embedded in L_{wc} and customized for different classes to minimize the intra-class distance, especially the tails.

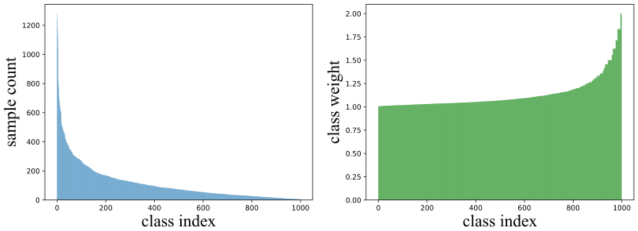

Figure 2: The long-tailed distribution of ImageNet-LT dataset and the corresponding weights of the wcenter loss.

Novelty Detection: LUNA

In the unfamiliarity detection procedure, we are aiming to localize each testing sample in the feature space with respect to the trained categories (acquaintance) to measure the level of novelty. As originally defined in the local outlier factor (Breunig et al. 2000), k-distance $d_k(p)$ is the distance between a feature vector p to its k-th nearest neighbor. Upon that, the reachability distance rd between the anchor p and another peer feature q is defined as the maximum of the usual distance between them and the k-distance of q, i.e.,

$$rd_k(p, q) = \max(d(p, q), d_k(q)). \tag{5}$$

since each sample is assigned to a cluster (or center) by its ground truth label for the training set, we define the *sub-local reachability density* (hereafter denoted as $\mathcal{D}$) for p, which is formulated as the inverse of the average of the k-distance between p to its k nearest neighbors N_k in the same cluster, where k is empirically chosen as half the cluster size.

$$\mathcal{D}_k(p) = 1 \Bigg/ \left(\frac{\sum_{q\in N_k} rd_k(p, q)}{|N_k(p)|}\right) \tag{6}$$

A smaller value of $\mathcal{D}$, which indicates a large distance between the anchor point to its neighbors, means the anchor is in a less dense neighborhood.

Inside each cluster, the points are classified into core if the $\mathcal{D}$ is within top η% of the points in the same cluster or boundary if not. Besides, we define the second boundary points as the points which belong to other clusters but regard this cluster as their second-best choice by the distance to its center point. Note that neighbors of second boundary points (see the purple dashed region in Figure 1) are chosen

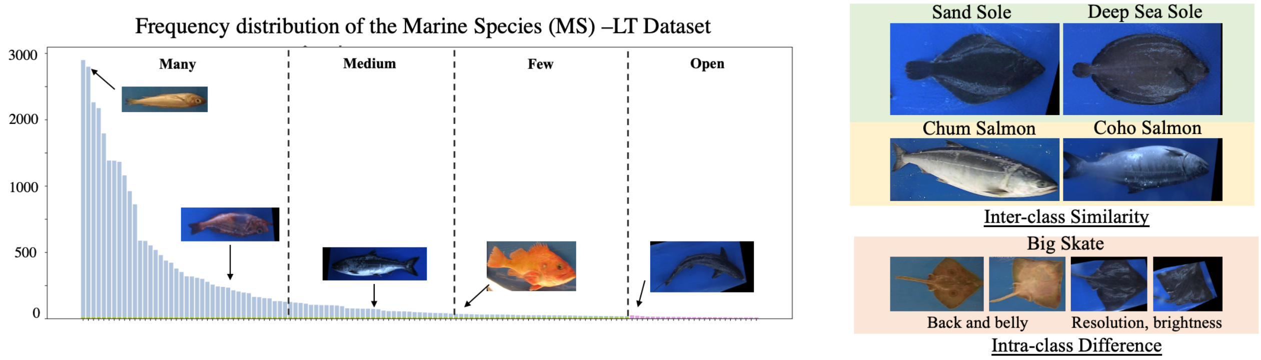

Figure 3: Left: the distribution of the proposed MS-LT dataset. There are three levels of frequency for the closed-set: many (counts > 100), medium ($20 <$ count ≤ 100) and few (count ≤ 20). The training set follows a long-tailed distribution, while the testing and validation sets are balanced following the configuration of other long-tailed datasets. Right: the challenging samples in the MS-LT. Some classes are similar in appearance, while some samples in the same class are different in orientation, resolution and lighting.

to be within this current cluster (see the yellow cluster in Figure 1), i.e., same as the core and boundary points, even though the second boundary points not necessarily belong to this cluster. The set of the core, boundary and second boundary points are denoted as $p_c \in S_C$, $p_b \in S_B$ and $p_{sb} \in S_{sB}$, respectively. By comparing a test point's $\mathcal{D}$ with those of the p_c, p_b and p_{sb}, we would be able to estimate its approximate location in the feature space. Theoretically, with greater density than $\mathcal{D}_{p_b}$ is in the safe zone to claim an inliner, while with sparser density than $\mathcal{D}_{p_{sb}}$ is likely to be a novel class. These probabilities are formulated by a set of outlier factors, i.e., core outlier factor (COF, $\mathcal{F}_C$), boundary outlier factor (BOF, $\mathcal{F}_B$) and second boundary outlier factor (sBOF, $\mathcal{F}_{sB}$), formulated as follows

$$
\begin{aligned}
\mathcal{F}_C(p) &= \frac{\sum_i^{S_C} \mathcal{D}_p / \mathcal{D}_{p_{c_i}}}{|S_C|}, \\
\mathcal{F}_B(p) &= \frac{\sum_i^{S_B} \mathcal{D}_p / \mathcal{D}_{p_{b_i}}}{|S_B|}, \\
\mathcal{F}_{sB}(p) &= \frac{\sum_i^{S_{sB}} \mathcal{D}_p / \mathcal{D}_{p_{sb_i}}}{|S_{sB}|}.
\end{aligned}
\tag{7}
$$

These metrics automatically adapt to the size of the clusters as they determine the number of nearest neighbors to be chosen. The LUNA factor $\mathcal{LF}$ of a deep feature p is thus defined as

$$
\begin{aligned}
\mathcal{LF}(p) = &\min\{|1 - \mathcal{F}_C(p)|, |1 - \mathcal{F}_B(p)|\} \\
&+ \left|1 - \frac{1}{|S_{sB}|} \sum_i^{S_{sB}} \mathcal{F}_{sB}(p_{sb_i}) / \mathcal{F}_{sB}(p)\right| + |1 - \theta_p|,
\end{aligned}
\tag{8}
$$

where the first term models p's density with respect to the core and boundary points in the same cluster; the second term is its density compared with the average $\mathcal{F}_{sB}$ of the second boundary points; and the last term θ_p is the maximum confidence output of the neural network. For testing samples in the open-set, its $\mathcal{F}_{sB}$ is close to or greater than the average $\mathcal{F}_{sB}$ of S_{sB}, leading to a smaller second term. For all three components, the larger the value, the more likely to be novel. Otsu's method is adopted to choose the optimal threshold.

Division	# of class	# of images
Many ($x > 100$)	43	23.3K
Medium ($20 < x \leq 100$)	32	1.7K
Few ($x \leq 20$)	31	0.4K
Total (close-set)	106	25.4K
Open-set	25	0.4K

Table 1: Statistics of our MS-LT dataset. Here, x denotes the number of samples in the class.

Dataset and Experiments

OLTR Datasets

The **ImageNet-LT** (Liu et al. 2019) dataset is re-sampled from a subset of the original ImageNet-2012 (Deng et al. 2009) following Pareto distribution. Extra 10 classes from ImageNet-2010 make up the open-set. There are 1000 classes for training, with 5 to 1280 images per class and 115.8K images in total.

The **Places-LT** (Liu et al. 2019) dataset is re-sampled from Place-2 dataset (Zhou et al. 2017) for scene recognition. There are 69 new classes in Place-Extra69 used as open-set.

Our proposed **Marine Species (MS)-LT** dataset is naturally long-tailed distributed (in Figure 3). It is collected from Gulf of Alaska and the Aleutian Islands in the U.S. during 2015 to 2019. There are 25.4K images for 106 marine species, with 5 to 1920 images per class. There are 25 classes for open-set, which were only observed in one of the years during collection. Table 1 shows the distribution and number of samples of MS-LT. The challenges are several classes share high inter-class similarity and some data of the same class exhibit vast differences in appearances with different orientations or are collected in different years.

The openness O (Bendale and Boult 2016) of an open-set is defined as

$$O = 1 - \sqrt{\frac{2 \times |C_{train}|}{|C_{test}| + |C_{target}|}}, \tag{9}$$

where C_{train} is the set of classes in training, C_{test} is the set of classes in testing and C_{target} is the set of classes to be identified. $|\cdot|$ denotes the size of the set. In the closed-set setting, the C_{target} is exactly the same as C_{test} and C_{train}, the openness is zero. The more the novel classes, the higher the openness, the more difficult the task. The openness of ImageNet-LT, Places-LT and MS-LT is 0.005, 0.085, and 0.331, respectively.

Evaluation Metrics

We first evaluate the Top-1 classification accuracy on the closed-set in the many, medium, and few splits. The splitting strategy, as mentioned in Table 1, is the same for all three datasets. After the open-set is merged in, we evaluate the OLTR performance by the Top-1 multi-class recognition accuracy on the close-set and open-set separately. Note that the close-set accuracy under the open-set setting is different from that in the closed-set setting, as some samples from known classes are also possible to be recognized as novel. Besides, following (Bendale and Boult 2016) and (Liu et al. 2019), we use the F-measure (or F-score), formulated as

$$\text{F-measure} = \frac{\text{TP}}{\text{TP} + 1/2(\text{FP} + \text{FN})}, \tag{10}$$

where the TP is the number of samples that are correctly predicted to their ground truth classes, while FP represents the known samples that are falsely classified but regarded within closed-set. FN is the number of known samples that falsely considered as the novel.

Implementation Details

The training samples are re-scaled by its shorter side and then resized to 224×224 with random crop and horizontal flip as data augmentation. We use ResNet-10, ResNet-152 and ResNet-32 as the backbone for ImageNet-LT, Place-LT and MS-LT, respectively. Following the two-stage decoupling training scheme (Kang et al. 2019), we first train the model with the original imbalanced dataset by stochastic gradient descent (SGD) with the momentum of 0.9 and weight decay of 2×10^{-4} in minibatch size of 128 for 180 epochs; then continue training the model with progressively-balanced re-sampling (Kang et al. 2019) with learning rate 0.05 for an extra 50 epochs. The wcenter loss is applied only at the second stage. There is no extra parameter to weigh different loss components.

Performance Comparison

Performance on public benchmarks. Following the baseline experiments in the OLTR network (Liu et al. 2019), we report the performance of the proposed method in the open-set and closed-set settings. The base model denotes the plain ResNet (He et al. 2016) without any adaptation on long-tailed or open-set configurations. Lifted loss (Oh Song et al. 2016), focal loss (Lin et al. 2017) and range loss (Zhang et al. 2017) are metric learning techniques to pull features of the same categories closer, where the range loss is designed for the long-tailed face recognition task. OpenMax (Bendale and Boult 2016) is a statistical fitting method to predict the probability of novelty in a similar manner as the SoftMax. FSLwF (Gidaris and Komodakis 2018) is a few-shot learning algorithm. OLTR (Liu et al. 2019) is the first work to formally define the OLTR problem and propose a network with visual memory and weight regularization to transfer knowledge from heads to tails as well as separate knowns and unknowns. IEM (Zhu and Yang 2020) designs region self-attention to improve the quality of memorized features.

Table 2 shows the performance on the ImageNet-LT, Places-LT and MS-LT, respectively. With our emphasis on long-tailed learning, our model outperforms the OLTR network by 3.4%, 1.5%, and 6.8%in overall accuracy of multi-class recognition (closed-set). It also improves the F-measure by 5.4%, 0.5%, and 5.4%, respectively. Our advantages lie in the improvements in the many and medium segments. We think it is critical to balance the heads and tails properly. Applying methods for the balanced set, like the base model, yields promising performance on the sample-rich categories but performs poorly on the tails. Another extreme attempt is to use a few-shot learning scheme, like FSLwF in the tables, to promote the tails' performance, but this is not advantageous in open-set testing. With various metric learning losses, such as the lifted loss, focal loss and range loss, the performance under the open-set setting is comparable with that of the closed-set. This supports our argument that representation learning is an effective tool for transferring the closed-set knowledge to applications of open-set. Therefore, we use input re-balancing and progressively adapt the classifier to handle the long-tailed problem. With the frequency-sensitive wcenter loss, the feature space is adequately organized and separable.

Discussions on the open-set performance. F-measure evaluates both the classification accuracy and novelty detection recall rate. However, the novelty detection does not make a significant difference if the open-set is small, i.e., small false negative value in Equation 10. The novelty detection accuracy, which is the portion of open-set being correctly identified, is a better metric to purely evaluate the model's ability of identifying the new classes. In Table 2, we report closed-set overall accuracy (false positives are the closed-set samples that are incorrectly classified as another closed-set class or open-set), open-set accuracy (false positives are the open-set samples that are misclassified as closed-set), and the F-measure. Comparing to the prior leading methods, LUNA achieves better performance on the novel classes, without sacrificing the long-tailed classification accuracy. The results show MS-LT is a challenging OSR dataset due to its high openness and fine-grained properties.

Visualize the feature space. Figure 4 shows the t-SNE (Maaten and Hinton 2008) visualization of the MS-LT dataset. With the wcenter loss, it is observed that the intra-class distances are reduced and each class is nicely clustered together, especially for the tails. The more concentrated they are, the more precise for LUNA factor estimation. The open-set samples are spread over empty region of the feature space

Dataset	Model	Closed-set				Open-set			
		Many	Medium	Few	Overall	Many	Medium	Few	F-measure
ImageNet-LT (ResNet-10)	Base Model	40.9	10.7	0.4	20.9	40.1	10.4	0.4	0.295
	Lifted Loss	35.8	30.4	17.9	30.8	34.8	29.3	17.4	0.374
	Focal Loss	36.4	29.9	16.0	30.5	35.7	29.3	15.6	0.371
	Range Loss	35.8	30.3	17.6	30.7	34.7	29.4	17.2	0.373
	OpenMax	-	-	-	-	35.8	30.3	17.6	0.368
	FSLwF	40.9	22.1	15.0	28.4	40.8	21.7	14.5	0.347
	OLTR	43.2	35.1	18.5	35.6	41.9	33.9	17.4	0.474
	IEM	48.9	44.0	24.4	43.2	46.1	42.3	20.1	0.525
	LUNA (Ours)	**51.8**	**48.6**	**26.2**	**46.6**	**48.2**	**44.7**	**23.6**	**0.579**
Places-LT (ResNet-152)	Base Model	45.9	22.4	0.4	27.2	45.9	22.4	0.4	0.366
	Lifted Loss	41.1	35.4	24	35.2	41.0	35.2	23.8	0.459
	Focal Loss	41.1	34.8	22.4	34.6	41.0	34.8	22.3	0.453
	Range Loss	41.1	35.4	23.2	35.1	41.0	35.3	23.1	0.457
	OpenMax	-	-	-	-	41.1	35.4	23.2	0.458
	FSLwF	43.9	29.9	29.5	34.9	38.1	19.5	14.8	0.375
	OLTR	44.7	37	25.3	35.9	44.6	36.8	25.2	0.464
	IEM	46.8	39.2	28.0	39.7	**48.8**	**42.4**	28.9	0.486
	LUNA (Ours)	**48.7**	**42.4**	**30.2**	**42.1**	48.1	41.6	**29.0**	**0.491**
MS-LT (ResNet-32)	Base Model	56.1	35.1	8.0	35.7	56.1	35.1	11.4	0.537
	Lifted Loss	53.2	42.3	12.6	38.0	53.0	42.2	12.4	0.549
	Focal Loss	57.3	44.6	18.5	42.1	57.0	42.8	15.4	0.576
	Range Loss	55.8	43.8	15.7	40.5	55.8	43.6	15.6	0.575
	OpenMax	-	-	-	-	54.2	44.9	12.8	0.564
	OLTR	57.8	49.8	28.6	46.8	56.7	45.3	23.6	0.603
	LUNA (Ours)	**61.2**	**56.6**	**34.6**	**52.0**	**60.4**	**51.8**	**30.4**	**0.657**

Table 2: OLTR performance of top-1 accuracy on the ImageNet-LT, Places-LT, and MS-LT datasets. Best results are marked in bold. Results on MS-LT with various backbones are discussed in the supplementary material.

rather than inside the clusters. Therefore, wcenter loss significantly benefits open-set detection.

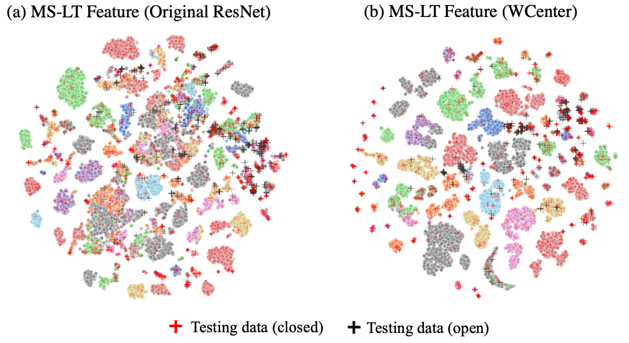

Figure 4: t-SNE visualization of the MS-LT dataset. Left: the original model; right: model with wcenter loss. The classes of training samples (dots) are marked in different colors. The testing set (closed) and open set are marked in red and black crosses, respectively.

Visualize the LUNA components. Figure 5 is the visualization of COF, BOF, sBOF and classification confidence of each sample in the MS-LT dataset. The COF and BOF of known classes are usually around 1, meaning they are close to the neighbors inside the cluster. They also have large sBOF as they have higher local density than the second boundary group. As for the novel points, the COF and BOF are smaller than 1 while the sBOF is close to 1, indicating they are far away from the cluster center like they are in another cluster. Besides, the network's outputs of maximum confidence, which is included in LUNA as well, are high for known classes and unstable for the novel classes.

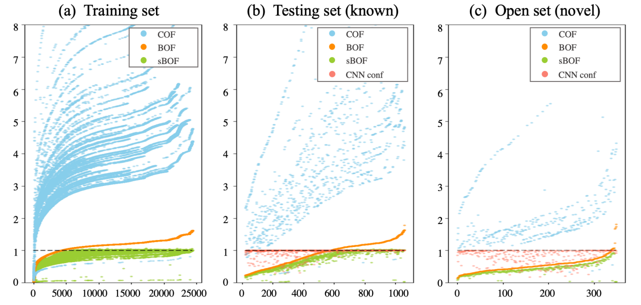

Figure 5: Visualization of COF, BOF, sBOF and confidence. The x-axis is the data point index, which is independent of each other. They are ordered by the value of BOF to show the trend. The black dash, $y = 1$, indicates the location of the point in the feature space. For example, a sample with COF near 1 is likely to be a core point of seen classes.

LOF and LUNA. To compare the novelty detection of LOF and LUNA, we use $K = 5$ for LOF, which is the size of the minimum cluster. The F-measure of LOF is 0.357 while the LUNA is 0.657. The reason is LOF selects a constant number of neighbors over the whole dataset for each testing sample, regardless of the clustering size or its potential category. LUNA uses variable sizes of neighbors that are adaptive to the clusters' sizes and different regions.

Ablation Study

Effect of wcenter loss. The role of wcenter loss is in two aspects: (1) re-weighting on the minority classes to benefit long-tailed recognition; (2) concentrating clusters in the feature space for outlier detection. Therefore, we compare the following schemes to show its effectiveness. Denoted λ_j as the weight of class j, which is a function of its frequency $\widetilde{n_j}$.

(a) None: $\lambda_j = 0$, which is training without center loss.

(b) Center: $\lambda_j = 1$, which is the vanilla center loss.

(c) Same: $\lambda_j = 1/\widetilde{n_j}$, which is the same as the frequency distribution.

(d) Inverse: $\lambda_j = \frac{\widetilde{n_j}}{\max_c\{\widetilde{n_c}\}}$, which is the inverse of the frequency distribution.

(e) Wcenter: $\lambda_j = \frac{\widetilde{n_j}}{\max_c\{\widetilde{n_c}\}} + 1$

The results, as shown in Table 3 are reported on MS-LT under the open-set setting. The result indicates bias on the feature domain affects the classification accuracy proportionally. The inverse loss and wcenter loss weight more on the tails, thus have more gains on the few-shot split. However, they sacrifice the accuracy on heads. Wcenter loss with a scaling term preserves the performance on head relatively. On the other hand, center loss and wcenter loss emphasis the clustering requirement, resulting in more desirable open-set detection performance. The result suggests metric learning is the key to solve open-set recognition problem with long-tailed training data: it is capable of helping imbalanced classification and automates feature selection in high dimension space for open-set recognition.

Weight	Many	Medium	Few	Overall	F-measure
None	62.4	49.8	29.4	48.8	0.632
Center	60.8	54.2	31.9	50.4	0.651
Same	63.5	55.8	28.4	50.8	0.608
Inverse	57.6	57.0	35.2	50.9	0.614
Wcenter	61.2	56.6	34.6	52.0	0.657

Table 3: Ablation study on weighting schemes on MS-LT. Many-/medium-/few-shot and overall accuracy are reported in closed-set; F-measurement is under open-set setting.

Effect of LUNA. LUNA measures the relative density of the testing sample with respect to the densities of core, boundary and second boundary in each cluster, as well as the network confidence. In this section, we show the experimental results by removing each component in Equation 8. To simplify, the three components are represented with its most important metric, i.e., $\mathcal{F}_C, \mathcal{F}_B, \mathcal{F}_{sB}$ and θ, respectively. The open-set performance on MS-LT is shown in Table 4.

The result indicates all the components in LUNA are necessary and effective. The first term evaluates the density regarding the core and boundary samples (inliers), which is shown to separate the majority from the novel samples. Removing it causes misclassification of the many-shot split. The second term and third term are responsible for separating the minority classes from the novel ones, as the clusters of minority classes are not as concentrated as the majorities', relaxing the metrics to the second boundary is beneficial.

Existing open-set recognition methods rely on the classification confidence (or the classifier output logits), which do not work well on long-tailed dataset as they do in balanced sets. Figure 5 (b) shows that confidence is not always high for closed-set samples, and it is the least important one comparing to the $\mathcal{F}_C, \mathcal{F}_B$ and $\mathcal{F}_{sB}$. Utilizing the sample's property itself (confidence), and the difference compared to the nearby acquaintance (trained samples) is more stable and interpretable.

Component	Many	Medium	Few	F-measure
LUNA	**60.4**	**51.8**	**30.4**	**0.657**
$-\mathcal{F}_C, \mathcal{F}_B$	54.5	46.8	25.1	0.607
$-\mathcal{F}_{sB}$	57.8	48.2	23.6	0.620
$-\theta$	59.2	48.4	27.4	0.636

Table 4: Ablation study on each component of LUNA.

Sensitivity of the hyper-parameter. We conduct experiments on the potion of each cluster (η) that is selected as core samples; results are shown in Table 5. As a joint evaluation of the density of multiple well-defined sample groups, the proposed LUNA is very robust in OLTR task and not sensitive to the selection of η. From our own test, we would recommend the η value from 0.2 to 0.5.

η	0.2	0.3	0.4	0.5	0.6	0.8
F-measure	0.654	0.657	0.655	0.648	0.635	0.622

Table 5: Ablation study on the potion of core samples.

Conclusions

In this research, we achieved "killing three birds with one stone". By introducing a fine-grained natural OLTR dataset about ocean fish species, researchers can engage to the real OLTR challenges in lab. Such a dataset can be a solid supplement to the existing, manually re-sampled OLTR benchmarks. Secondly, a new wcenter loss is designed to minimize the intra-class distance in the feature space, which preserves classification accuracy while optimizing the clustering for both the heads and tails. In addition, we propose the LUNA, which is an effective measure of novelty based on the relative local density of the learned representation. Our proposed LUNA significantly outperforms the SOTA OLTR algorithms in all three datasets.

References

1998. THE MNIST DATABASE. http://yann.lecun.com/exdb/mnist/. Accessed: 2021-09-06.

2019. Summary of Notifiable Diseases-Centers for Disease Control and Prevention, USA. https://www.cdc.gov/mmwr/mmwr_nd/index.html. Accessed: 2021-09-06.

Bendale, A.; and Boult, T. E. 2016. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1563–1572.

Breunig, M. M.; Kriegel, H.-P.; Ng, R. T.; and Sander, J. 2000. LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 93–104.

Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; and Ma, T. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, 1567–1578.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357.

Chopra, S.; Hadsell, R.; and LeCun, Y. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, 539–546. IEEE.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-Balanced Loss Based on Effective Number of Samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4690–4699.

Galar, M.; Fernández, A.; Barrenechea, E.; and Herrera, F. 2013. EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. *Pattern recognition*, 46(12): 3460–3471.

Ge, Z.; Demyanov, S.; Chen, Z.; and Garnavi, R. 2017. Generative openmax for multi-class open set classification. *arXiv preprint arXiv:1707.07418*.

Gidaris, S.; and Komodakis, N. 2018. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4367–4375.

Han, H.; Wang, W.-Y.; and Mao, B.-H. 2005. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, 878–887. Springer.

Hassen, M.; and Chan, P. K. 2020. Learning a neural-network-based representation for open set recognition. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, 154–162. SIAM.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, C.; Li, Y.; Loy, C. C.; and Tang, X. 2016. Learning Deep Representation for Imbalanced Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Júnior, P. R. M.; De Souza, R. M.; Werneck, R. d. O.; Stein, B. V.; Pazinato, D. V.; de Almeida, W. R.; Penatti, O. A.; Torres, R. d. S.; and Rocha, A. 2017. Nearest neighbors distance ratio open-set classifier. *Machine Learning*, 106(3): 359–386.

Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; and Kalantidis, Y. 2019. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*.

Kriegel, H.-P.; Kröger, P.; Schubert, E.; and Zimek, A. 2009. LoOP: local outlier probabilities. In *Proceedings of the 18th ACM conference on Information and knowledge management*, 1649–1652.

Kulis, B.; et al. 2012. Metric learning: A survey. *Foundations and trends in machine learning*, 5(4): 287–364.

Li, Y.; Wang, T.; Kang, B.; Tang, S.; Wang, C.; Li, J.; and Feng, J. 2020. Overcoming Classifier Imbalance for Long-Tail Object Detection With Balanced Group Softmax. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollar, P. 2017. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Liu, X.-Y.; Wu, J.; and Zhou, Z.-H. 2008. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2): 539–550.

Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; and Yu, S. X. 2019. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2537–2546.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov): 2579–2605.

Narayanan, A.; Dwivedi, I.; and Dariush, B. 2019. Dynamic traffic scene classification with space-time coherence. In *2019 International Conference on Robotics and Automation (ICRA)*, 5629–5635. IEEE.

Nguyen, H. M.; Cooper, E. W.; and Kamei, K. 2011. Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms*, 3(1): 4–21.

Oh Song, H.; Xiang, Y.; Jegelka, S.; and Savarese, S. 2016. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4004–4012.

Scheirer, W. J.; Jain, L. P.; and Boult, T. E. 2014. Probability models for open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 36(11): 2317–2324.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Tu, B.; Zhou, C.; Kuang, W.; Guo, L.; and Ou, X. 2018. Hyperspectral imagery noisy label detection by spectral angle local outlier factor. *IEEE Geoscience and Remote Sensing Letters*, 15(9): 1417–1421.

Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8769–8778.

Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; and Liu, W. 2018. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5265–5274.

Wang, W.; and Lu, P. 2011. An efficient switching median filter based on local outlier factor. *IEEE Signal Processing Letters*, 18(10): 551–554.

Wang, X.; Deng, X.; Fu, Q.; Zhou, Q.; Feng, J.; Ma, H.; Liu, W.; and Zheng, C. 2020a. A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT. *IEEE transactions on medical imaging*, 39(8): 2615–2625.

Wang, X.; Lian, L.; Miao, Z.; Liu, Z.; and Yu, S. X. 2020b. Long-tailed Recognition by Routing Diverse Distribution-Aware Experts. *arXiv preprint arXiv:2010.01809*.

Wellmer, F.-W.; and Becker-Platen, J. D. 2000. Global nonfuel mineral resources and sustainability. In *Proceedings for a Workshop on Deposit Modeling, Mineral Resource Assessment, and Their Role in Sustainable Development*, 1.

Wen, Y.; Zhang, K.; Li, Z.; and Qiao, Y. 2016. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, 499–515. Springer.

Zhang, X.; Fang, Z.; Wen, Y.; Li, Z.; and Qiao, Y. 2017. Range loss for deep face recognition with long-tailed training data. In *Proceedings of the IEEE International Conference on Computer Vision*, 5409–5418.

Zhou, B.; Cui, Q.; Wei, X.-S.; and Chen, Z.-M. 2020. BBN: Bilateral-Branch Network with Cumulative Learning for Long-Tailed Visual Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9719–9728.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6): 1452–1464.

Zhu, J.; Wang, Y.; Zhou, D.; and Gao, F. 2018. Batch process modeling and monitoring with local outlier factor. *IEEE Transactions on Control Systems Technology*, 27(4): 1552–1565.

Zhu, L.; and Yang, Y. 2020. Inflated Episodic Memory With Region Self-Attention for Long-Tailed Visual Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4344–4353.