Open Long-Tailed Recognition In A Dynamic World

Ziwei Liu*, Zhongqi Miao*, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu.

Abstract—Real world data often exhibits a long-tailed and open-ended (*i.e.* with unseen classes) distribution. A practical recognition system must balance between majority (head) and minority (tail) classes, generalize across the distribution, and acknowledge novelty upon the instances of unseen classes (open classes). We define Open Long-Tailed Recognition++ (OLTR++) as learning from such naturally distributed data and optimizing for the classification accuracy over a balanced test set which includes both known and open classes. OLTR++ handles imbalanced classification, few-shot learning, open-set recognition, and active learning in one integrated algorithm, whereas existing classification approaches often focus only on one or two aspects and deliver poorly over the entire spectrum. The key challenges are: 1) how to share visual knowledge between head and tail classes, 2) how to reduce confusion between tail and open classes, and 3) how to actively explore open classes with learned knowledge. Our algorithm, OLTR++, maps images to a feature space such that visual concepts can relate to each other through a memory association mechanism and a learned metric (dynamic meta-embedding) that both respects the closed world classification of seen classes and acknowledges the novelty of open classes. Additionally, we propose an active learning scheme based on visual memory, which learns to recognize open classes in a data-efficient manner for future expansions. On three large-scale open long-tailed datasets we curated from ImageNet (object-centric), Places (scene-centric), and MS1M (face-centric) data, as well as three standard benchmarks (CIFAR-10-LT, CIFAR-100-LT, and iNaturalist-18), our approach, as a unified framework, consistently demonstrates competitive performance. Notably, our approach also shows strong potential for the active exploration of open classes and the fairness analysis of minority groups.

Index Terms—Long-Tailed Recognition, Few-shot Learning, Active Learning.

1 Introduction

Our visual world is inherently long-tailed and open-ended [1], with a few common visual categories (i.e., head classes) and many more relatively rare categories (i.e., tail classes). At the same time, new visual concepts constantly emerge as we navigate in an open world (i.e., open classes).

Although natural data distributions contain head, tail, and open classes, existing classification approaches focus mostly on either the head [2], [3] or the tail [4], [5], and often in a closed setting [6], [7] (Fig. 2). We thus formally study *Open Long-Tailed Recognition++* (OLTR++) arising from natural data settings. A practical system should be able to work for a few head and many tail categories, to generalize the concept of a single category from only a few known instances, as well as to acknowledge and explore novelty upon an instance of an unseen or open category. We define OLTR++ as learning from long-tail and open-end distributed data and evaluating the classification accuracy over a balanced test set which includes head, tail, and open classes in a continuous spectrum (Fig. 1).

The key challenges for OLTR++ are tail recognition robustness and open-set sensitivity. As the number of training instances drops from thousands in the head class to a few in the tail class, we should prevent the performance from dropping drastically. Meanwhile, for open classes, the recognition performance relies on the sensitivity to distinguish unknown samples from known classes, as well as to select informative samples for data-efficient active exploration and future model updates.

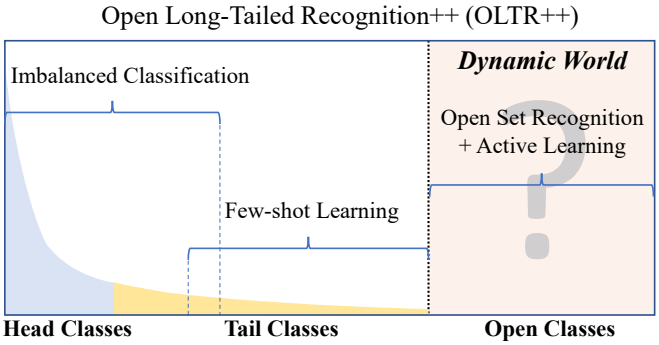


Fig. 1: **OLTR++.** Our task of open long-tailed recognition++ learns from long-tail distributed training data in an open world and deals with imbalanced classification, few-shot learning, open-set recognition, and active learning over the entire spectrum.

In our algorithm (OLTR++), there is a learned metric that respects both the closed-world classification and acknowledge the novelty of the open world. This metric maps images to an embedding space (*dynamic meta-embedding*), where visual concepts can relate to each other to improve both the tail recognition robustness and open-set sensitivity.

Specifically, our *dynamic meta-embedding* is a combination of two components: direct feature and induced feature. **1)** Direct feature is a standard embedding that gets updated from the training data by stochastic gradient descent over the classification loss. It is usually less generalized in tail classes, compared to head classes, because of the lack of sufficient supervision. **2)** Memory feature, instead, is an induced feature through a visual memory association mechanism, inspired by meta-learning methods [4],

- *Z. Liu is with Nanyang Technological University. Z. Miao, J. Wang and S. Yu are with the University of California, Berkeley and International Computer Science Institute. X. Zhan is with The Chinese University of Hong Kong. B. Gong is with Google Inc.*
- **Equal contribution.*

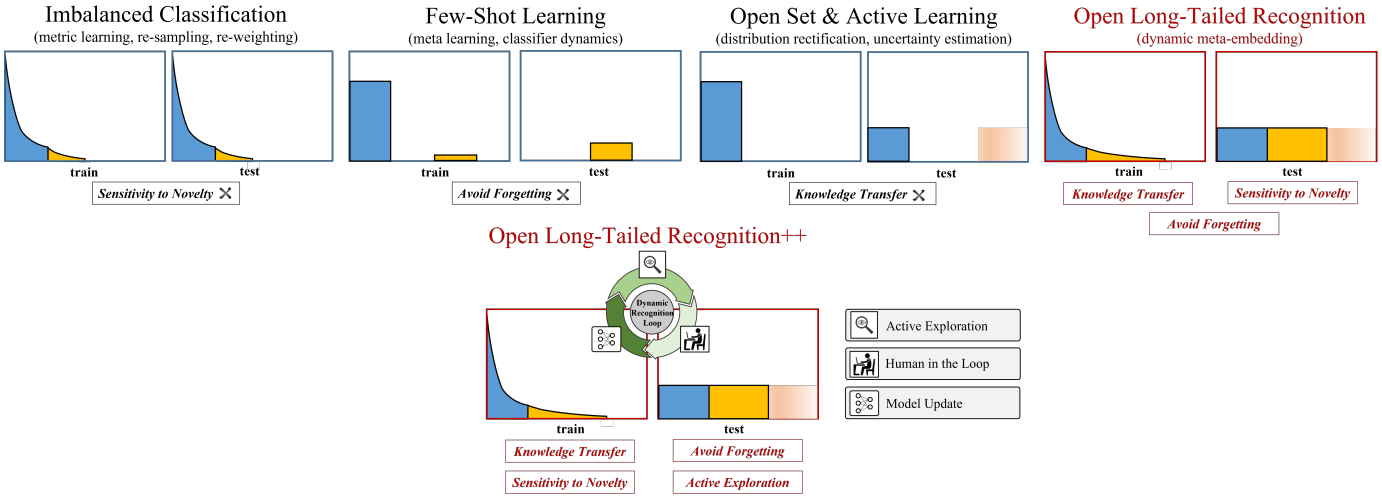


Fig. 2: **Comparison between our proposed OLTR++ and related existing tasks.** Our task, in a more realistic setting, is a combination of the three existing tasks (imbalanced classification, few-shot learning, and open-set recognition) in addition to active exploration, which constitutes a dynamic learning loop that can facilitate robust deployment of vision systems.

[8], [9], that augments the direct feature of each image for better distinguishability. The association is learned to retrieve a summary of memory activations from the direct features of each input and then combined with the original direct features to be the meta-embeddings. This memory feature augmentation is particularly effective on tail classes for the lack of supervision to provide generalized features.

We address open-class sensitivity by dynamically calibrating the meta-embedding with respect to the visual memory. In other words, the embedding is scaled inversely by its distance to the nearest class centroid (i.e., class memory): the farther away it is from the memory, the more likely it is an open set instance. The distance between the embedding and the nearest centroid is then transformed into "sample informativeness" using an energy-based model [10], which is further employed to select informative samples for active learning.

We also adopt *modulated attention* [11] to encourage the head and tail classes to use different sets of spatial features. As our meta-embedding relates head and tail classes, our modulated attention maintains discrimination between them.

We make the following major contributions. **1)** We formally define the OLTR++ task, which learns from long-tail and open-end distributed data and optimizes for the overall accuracy over a balanced test set. It provides a comprehensive and unbiased evaluation of visual recognition algorithms in practical settings. **2)** We propose an integrated OLTR++ algorithm with dynamic meta-embedding. It handles tail recognition robustness by relating visual concepts among head and tail embeddings, and it handles open-class sensitivity by dynamically calibrating the embedding with respect to the visual memory. **3)** We further incorporate an energy-based model into our dynamic meta-embedding for data-efficient active learning (with only sparse human annotations), which is well suited for the constantly-changing visual world. **4)** We curated three large OLTR++ datasets according to a long-tailed distribution from existing representative datasets: object-centric ImageNet [2], scene-centric Places [12], and face-centric MS1M datasets [13]. We also set up benchmarks for proper OLTR performance evaluation. **5)** Our extensive experimentation on these OLTR++ datasets (as well as standardized benchmarks such as CIFAR-LT 100/10 and iNaturalist-18 [14]) demonstrates that our method consistently outperforms the state-of-the-art methods.

The aim of this work is to advocate a new learning paradigm that can perceive and update in a dynamic world, *i.e.* simultaneously recognizing real-world long-tailed data while actively exploring novel data with human in the loop. It is a crucial step towards embodied intelligence as well as the robust deployment of vision systems. Besides exhibiting competitive performance on long-tailed recognition, our approach also demonstrates compelling results on open class detection and active exploration with a unified framework centered around visual memory.

Our code, datasets, and models are publicly available at https://liuziwei7.github.io/projects/LongTail.html. Our work fills the void in practical benchmarks for imbalanced classification, few-shot learning, open-set recognition, and active learning, enabling future research that is directly transferable to real-world applications.

2 Related Works

While OLTR++ has not been defined in the previous literature, there are four closely related tasks which are often studied in isolation: imbalanced classification, few-shot learning, open-set recognition, and active learning. Fig. 2 summarizes their connections and key differences.

Imbalanced & Long-Tailed Recognition. Imbalanced classification is an extensively studied area [23], [24], [25], [26], [27], [28], [29], [30], [31]. While classical methods include under-sampling head classes, over-sampling tail classes, and data instance re-weighting, some recent methods apply *metric learning* [32], [33], *hard negative mining* [16], [34], and *meta learning* [6], [35]. For example, lifted structure loss [33] introduces margins between many training instances. The range loss [36] enforces data in the same class to be close and those in different classes to be far apart. Focal loss [16] induces an online version of hard negative mining. And metaModelNet [6] learns a meta regression net from head classes and uses it to construct the classifier for tail classes.

As a step forward from imbalanced classification, long-tailed recognition (where the training datasets are more imbalanced) [17], [18], [19], [20], [37], [38] has attracted extensive

Method	Formulation	Venue	CIFAR-10	CIFAR-100	ImageNet-LT	Places-LT
Vanilla FC [15]	$W_i^T f(x)$	CVPR 16	70.4	38.3	20.9	27.2
Hard Example mining [16]	$r(y) = (1-p_y)^\gamma$	ICCV 17	70.4	38.4	30.5	34.6
Re-Weighting [17]	$r(y) = (1-\beta)/(1-\beta_y^n)$	CVPR 19	74.6	36.0	29.7	38.9
Memory **(ours)** [18]	$W_i^T f^{memory}(x)$	CVPR 19	76.3	41.2	39.6	39.3
Class-Aware Margin [19]	$W_i^T f(x) - \mathbb{1}\{i=y\} \cdot \delta_i$	NeurIPS 19	77.0	42.0	36.2	35.7
Classifier Re-Scaling [20]	$s_i \cdot W_i^T f(x)$	ICLR 20	-	43.2	41.4	36.7
Bilateral Branch [21]	$(W_i^{bilateral})^T f(x)$	CVPR 20	79.8	42.6	-	-
Multi Experts [22]	$(W_i^{expert})^T f(x)$	ICLR 21	-	47.0	54.4	-

TABLE 1: A systematic overview of representative long-tailed recognition works (including those published after ours). Suppose there are C classes in total with $n_i, i \in \{1, 2, ..., C\}$ samples for class i. We denote $f(x)$ as the deep feature extracted from image x and $W = [W_1, ..., W_C]$ as the classifier weight vectors. The accuracies are reproduced with their released code.

Problem	Known Classes	Unknown Classes
Active Learning	✓ (informativeness)	
Active Exploration	✓ (informativeness)	✓ (info. & openness)

TABLE 2: **Key differences between active learning and active exploration.** "info." stands for informativeness.

research interests recently. In Table 1, we provide a systematic overview of representative long-tailed recognition works (including those published after ours).

In our approach, we have a dynamic meta-embedding that combines the strengths of both metric learning and meta learning. On the one hand, our direct feature is updated to ensure centroids of different classes are separated from each other; On the other hand, our memory feature is generated on-the-fly in a meta learning fashion to effectively transfer knowledge from head classes to tail classes.

Few-Shot Learning. Few-shot learning is often addressed by meta-learning techniques [39], [40], [41], [42], [43], [44]. For example, matching Network [4] learns a transferable feature matching metric to go beyond given classes. And Prototypical Network [45] maintains a set of separable class templates. Additionally, feature hallucination [46] and augmentation [47] have also shown effectiveness. Since these methods focus on fast learning from novel and unseen classes, they often suffer a "catastrophic forgetting" for training classes . Few-shot learning without forgetting [48] and incremental few-shot learning [49] attempt to remedy this issue by leveraging the duality between features and classifiers' weights [50], [51]. However, these methods rely on balanced training sets, while OLTR++ learns from naturally long-tailed training sets instead.

Our approach is closely related to meta learning with associative memory [4], [8], [9], [52], [53], [54]. Compared to prior arts, our memory feature has two advantages: **1)** it transfers knowledge to both head and tail classes adaptively via a learned concept selector; **2)** it is fully integrated into the network without episodic training, thus suitable for large-scale applications.

Open-Set Recognition. Open-set recognition [55], [56], or out-of-distribution detection [57], [58], [59], aims to re-calibrate the sample confidence in the presence of open classes. One of the representative techniques is OpenMax [56], which fits a Weibull distribution to model logits. However, when there are both open and tail classes, distribution-fitting based methods often confuse the two because of the less generalized features of tail classes. Instead of calibrating the output logits, our OLTR++ approach incorporates a confidence estimation mechanism into feature learning and dynamically re-scale the meta-embedding w.r.t. the learned visual memory, such that samples from known classes are expected to be closer to the memory compared to novel samples.

Open-World Recognition. Open-world recognition [60], [61], [62] is a closely related field whose goal is to distinguish "unknown unknown classes" from "known known classes". [60] also considers a dynamic setting where unknown classes are continuously added and detected, and examines the influence of unknown classes on the accuracy of known classes. [62] further incrementally learns the new classes. Once they are detected as unknown and an oracle provides labels for the objects of interest among all the unknowns. Here we advocate this dynamic-world endeavor: instead of just detecting the unknown classes, we aim to recognize the semantic label of the unknown classes.

Zero-Shot Learning. Zero-shot learning (ZSL) [63], [64], [65] is also a promising direction for recognizing novel classes. ZSL aims to learn the association between base and novel class features with the aid of certain shared semantic knowledge (*e.g.* attributes, word2vec), which is not directly applicable here. In comparison, our active exploration is more focused on the annotation-efficiency of recognizing novel classes, *i.e.* using less human annotations to achieve acceptable accuracies on novel classes.

Active Learning. Active learning aims to explore unlabeled data with an oracle annotator that provides ground truth labels to a few selected samples. The central issue here is the exploration efficiency, *i.e.* obtaining higher performance with less oracle queries. The representative works can be roughly categorized into two realms: generation-based methods [66], [67] and selection-based methods [68], [69].

However, existing active learning methods mainly work in closed-world setting, where they focus on selecting informative samples for the known categories to improve the performance. Here we study a realistic yet more challenging problem, selecting informative samples from a mixture of known and unknown categories so as to recognize both of them, which we coined as "active exploration". Their key differences are listed in Table 2. Our active exploration considers both informativeness and openness during sample selection.

Incremental Learning. Incremental learning aims to continually learn new tasks (*e.g.* novel classes) without catastrophic forgetting. Extensive research has been performed from different perspectives: neural architectures [70], [71], replay/rehearsal mechanisms [72], [73], parameter regularization [74], [75] and learning techniques [76], [77]. Here our dynamic learning loop

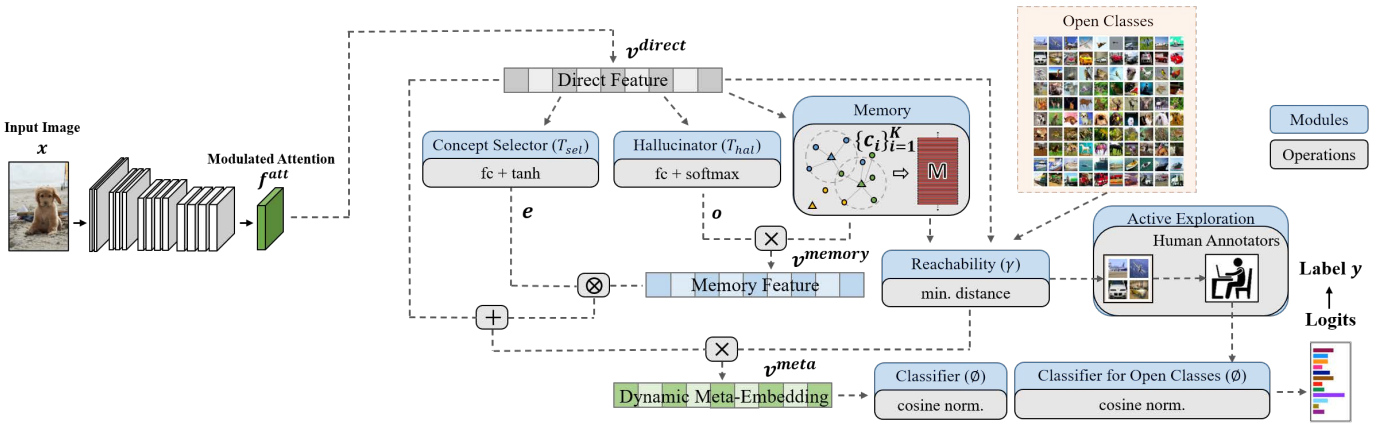


Fig. 3: **OLTR++ framework overview.** There are two main modules: dynamic meta-embedding and modulated attention. The embedding relates visual concepts between head and tail classes, while the attention discriminates between them. The *r*eachability separates known and open classes.

Problem	Imbalanced Asp.	Optimization Obj.
Fairness Analysis	sensitive attributes	attribute-wise criteria
Long-Tailed Recognition	categories	acc. on all categories

TABLE 3: **Key differences between fairness analysis and open long-tailed recognition.** "asp." stands for aspects while "obj." stands for objectives.

aims to achieve continual recognition of novel classes in a data-efficient manner.

Fairness Learning/Analysis. The open long-tailed recognition proposed in our work also has an intrinsic relationship to fairness analysis [78], [79], [80], [81], [82]. Their key differences are listed in Table 3. On the problem setting side, both open long-tail recognition and fairness analysis aim to tackle the imbalance existed in real-world data. Open long-tailed recognition focuses on the longtail-ness in both known and unknown categories while fairness analysis deals with the bias in sensitive attributes such as male/female and white/black.

On the methodology side, both open long-tailed recognition and fairness analysis aim to learn transferable representations. Open long-tailed recognition optimizes for the overall accuracy of all categories while fairness analysis optimizes for several attribute-wise criteria. The preliminary results in Table 6 demonstrates that our proposed dynamic meta-embedding is also a promising solution to fairness analysis.

3 Our OLTR++ Model

We propose to map an image to a feature space such that visual concepts can relate to each other based on a learned metric that respects the closed-world classification while acknowledging the novelty of the open and dynamic world. Our model has three major modules (Fig. 3): *dynamic meta-embedding*, *modulated attention*, and *active exploration*. The first relates and transfers knowledge between head and tail classes, and the last two maintain discrimination between them with human-in-the-loop.

3.1 Intuitive Explanation of Our Approach

In this section, we give an intuitive explanation of our approach that tackles the problem open long-tailed recognition. From the perspective of knowledge gained from observation (*i.e.* training set), head classes, tail classes and open classes form a continuous spectrum as illustrated in Fig. 4.

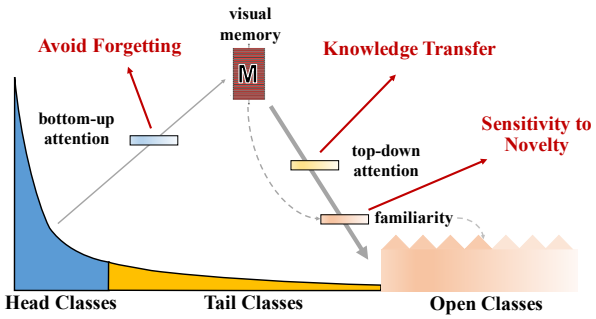


Fig. 4: **Intuition explanation** of our approach.

Firstly, in our approach, we obtain a *visual memory* by aggregating the knowledge from both head and tail classes. Then the visual concepts stored in the memory are infused back as associated "fast feature" to enhance the original "slow feature". In other words, we use induced knowledge (*i.e.* "fast feature") to assist the direct observation (*i.e.* "slow feature"). We further learn a *concept selector* to control the amount and type of infused "fast feature". Since head classes already have abundant direct observation, only a small amount of "fast feature" is needed. On the contrary, tail classes suffer from scarce observation, the associated visual concepts in "fast feature" are more beneficial to tail classes than to head classes. Finally, we calibrate the confidence of open classes by calculating their *reachabilities* (*i.e.* feature space distances) to the obtained visual memory (i.e., class centroids). All together, we provide a comprehensive treatment to the full spectrum of head, tail and open classes, improving the performance on all categories.

3.2 Dynamic Meta-Embedding

Dynamic meta-embedding is a combination of a direct image feature and an associated memory feature, where the feature norm indicates the familiarity to known classes.

Consider a convolutional neural network (CNN) with a softmax output layer for classification. The second-to-the-last

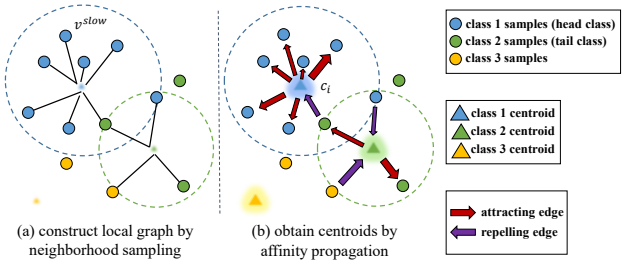


Fig. 5: **The discriminative centroids** constitute our visual memory, which are obtained with two iterative steps, neighborhood sampling and affinity propagation.

layer is the feature and the last layer is the linear classifier (cf. $\phi(\cdot)$ in Fig. 3). The feature and the classifier are jointly trained in an end-to-end fashion. Let v^{direct} denote the *direct feature* extracted from an input image. The final classification accuracy largely depends on the quality of this direct feature.

While a feed-forward CNN classifier works well with big training data [2], [83], it lacks sufficient supervised updates from small data in tail classes. We propose to enrich direct feature v^{direct} with visual concepts in a memory module through a memory feature v^{memory}, which is derived from a memory bank that captures visual concepts of each training classes. This mechanism is similar to the memory components in meta learning [42], [54]. We denote the resulting feature *meta embedding* v^{meta}.

Learning Visual Memory M**.** To construct the memory bank, we follow [84], [85] on class structure analysis and adopt discriminative centroids as the basic building block. Let M denote the visual memory of all the training data, $M = \{c_i\}_{i=1}^{K}$ where K is the number of training classes. Compared to alternatives [45], [86], this memory is appealing for our OLTR++ task: it is almost effortlessly and jointly learned alongside the direct features $\{v_n^{direct}\}$, and it considers both intra-class compactness and inter-class discriminativeness.

More concretely, as illustrated in Fig. 5, we compute class centroids in two steps.

1) **Neighborhood Sampling:** We sample both intra-class and inter-class examples to compose a mini-batch during training. These examples are grouped by their class labels and the centroid c_i of each group is updated by the direct feature of this mini-batch, which can be formulated as:

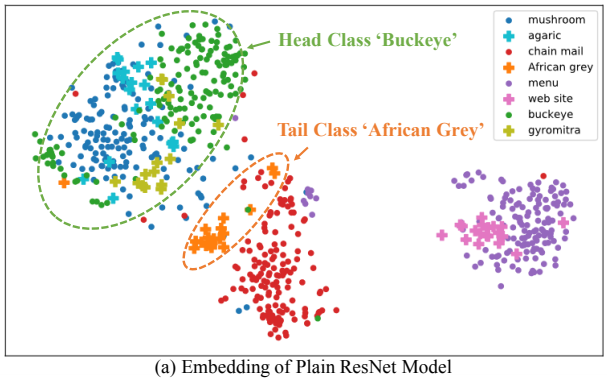
$$\Delta c_i = \frac{\sum_{b=1}^{B} \mathbb{1}(y_b = i) \cdot (c_i - x_b)}{1 + \sum_{b=1}^{B} \mathbb{1}(y_b = i)}, \tag{1}$$

where $\mathbb{1}(\cdot)$ is the indicator function, and B is the batch size.

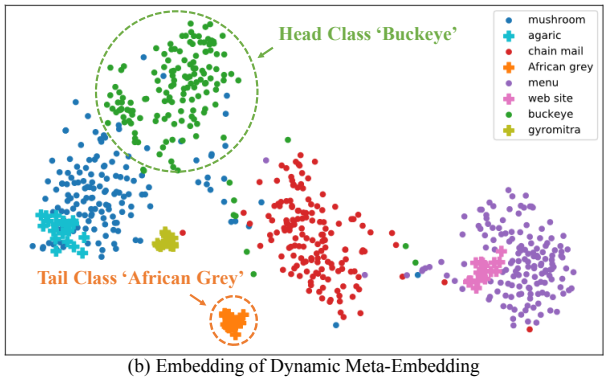
2) **Affinity Propagation:** We alternatively update the direct feature v^{direct} and the centroids to minimize the distance between each direct feature and the corresponding class centroids and maximize the distance to other centroids. Note that the "repelling edges" in Fig. 5 are calculated through a large margin loss L_{LM} as described in Eqn. 13.

At the end of the training, we obtain a visual memory module M containing important visual concepts within the dataset.

Composing Memory Feature v^{memory}**.** For an input image, v^{memory} enhances its direct feature when training data are limited



(a) Embedding of Plain ResNet Model



(b) Embedding of Dynamic Meta-Embedding

Fig. 6: **t-SNE feature visualization** of (a) plain ResNet model (b) our *dynamic meta-embedding*. Ours is more compact for both head and tail classes.

(as in the tail class). The memory feature relates class centroids in the memory to transfer knowledge to tail classes:

$$v^{memory} = o^T M := \sum_{i=1}^{K} o_i c_i, \tag{2}$$

where $o \in \mathbb{R}^K$ is the coefficients hallucinated from the direct feature. We use a lightweight neural network to obtain the coefficients from the direct feature, $o = T_{hal}(v^{direct})$.

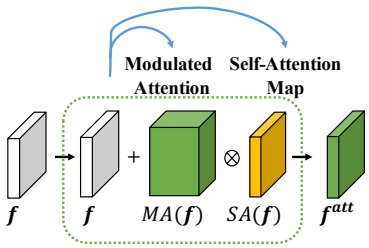
Obtaining Dynamic Meta-Embedding. v^{meta} combines the direct feature and the memory feature, and is fed to the classifier for the final class prediction (Fig. 6):

$$v^{meta} = (1/\gamma) \cdot (v^{direct} + e \otimes v^{memory}), \tag{3}$$

where $\otimes$ denotes element-wise multiplication and e is a concept selector. $\gamma > 0$ is a seemingly redundant scalar for the closed-world classification tasks. However, in the OLTR++ setting, it serves an important role in differentiating the examples of the training classes from those of the open-set. γ measures the reachability [87] of an input's direct feature v^{direct} to the memory M — the minimum distance between the direct feature and the discriminative centroids:

$$\gamma := \text{reachability}(v^{direct}, M) = \min_i \|v^{direct} - c_i\|_2. \tag{4}$$

When γ is small, the input likely belongs to a training class from which the centroids are derived, and a large reachability weight $1/\gamma$ is assigned to the resulting meta-embedding v^{meta}. Otherwise, the embedding is scaled down to an almost all-zero vector at the extreme, which is useful to encode open classes.



(a) Modulated Attention

(b.1) Input Image (b.2) Feature Map of Plain ResNet Model (b.3) Feature Map of Our Model (b.4) Modulate Attention

Fig. 7: **Modulated attention** is spatial attention applied on self-attention maps ("attention on attention"). It encourages different classes to use different contexts, which helps maintain the discrimination between head and tail classes.

As the direct feature is often good enough for the data-rich head classes, whereas the memory feature is more important for the data-poor tail classes. We design a concept selector, e, to adaptively select them in a soft manner. We learn a lightweight network $T_{sel}(\cdot)$ with a $\tanh(\cdot)$ activation function on v^{direct}:

$$e = \tanh(T_{sel}(v^{direct})). \tag{5}$$

3.3 Modulated Attention

Since discriminative cues of head and tail classes tend to distribute at different locations in the image, we find it beneficial to enhance direct feature, v^{direct}, with spatial attention to separate head and tail classes. Specifically, we propose *modulated attention* to encourage samples of different classes to use different contexts. Firstly, we compute a self-attention map $SA(f)$ from the input feature map by self-correlation [11]. It is used as contextual information and added back (through skip connections) to the original feature map. The *modulated attention* $MA(f)$ is then designed as conditional spatial attention applied to the self-attention map: $MA(f) \otimes SA(f)$, which allows examples to select different spatial contexts (Fig. 7). The final attention feature map becomes:

$$f^{att} = f + MA(f) \otimes SA(f), \tag{6}$$

where f is a feature map in CNN, $SA(\cdot)$ is the self-attention operation, and $MA(\cdot)$ is a conditional attention function [88] with a softmax normalization.

Sec. 4.1 shows empirically that our attention design achieves superior performance than the common practice of applying spatial attention to the input feature map. This modulated attention (Fig. 7 (b)) could be plugged into any feature layer of a CNN. Here, we modify the last feature map only.

3.4 Active Exploration of Open Classes

In the dynamic world, the model should not halt after training. We assume a continuous training, inference, annotation, and model update loop as our model actively explores the visual world over time. Every time our model encounters certain sample of open classes, our model will determine whether this sample is informative enough for further human annotation. After obtaining these human annotations in an efficient manner, our model will be continually updated according to the newly added data.

Algorithm 1 Active Exploration of Open Classes.

Input:
v^{direct}: the direct feature extracted from the open sample,
c_i: the discriminative centroid of class i from visual memory,
K: the number of classes,
T_{act}: the temperature for trade-off in active exploration.
for each exploration step **do**
  Sample mini-batch $\{v_n^{direct}\}$.
  Compute the minimum distance between the direct feature and the discriminative centroid:
  $U_{open} \leftarrow \min_i \|v^{direct} - c_i\|_2$.
  Compute the ratio between the first two nearest distances:
  $U_{info} \leftarrow d_1^{sorted} \,/\, d_2^{sorted}$.
  Compute the free energy function of v^{direct}:
  $E_n \leftarrow -T_{act} \cdot \log \sum_i^K e^{U_{open} \cdot U_{info}/T_{act}}$.
  Select high-energy samples for further human annotations.
  Update the classifier $\phi(\cdot)$ using newly added data.
end for

The active exploration step has three major components: 1) active sample selection based on two different types of uncertainty, 2) human-in-the-loop annotation, and 3) model update using active data annotations, all three of which constitute a dynamic recognition loop. The detailed algorithmic pipeline of our active exploration is listed in Alg. 1.

3.4.1 Two Types of Uncertainty in Active Exploration

Unlike the standard active learning setting that work in closed-world setting, there actually exist two types of uncertainty here in active exploration: **uncertainty in openness** and **uncertainty in informativeness**. Existing active learning algorithms are not directly applicable here since their uncertainty estimation mechanism only considers the informativeness among known classes, which is not suitable for modeling the openness between known classes and unknown classes.

In the following, we elaborate the modeling the two types of uncertainty in the context of active exploration:

1) **Uncertainty in Openness:** We measure the openness U_{open} of a new sample using the distance between its embedding and the nearest centroid, which can be formulated as:

$$U_{open} = \min_i \|v^{direct} - c_i\|_2, \tag{7}$$

where v^{direct} is the direct feature of the new sample and c_i is the centroid of the i-th class.

2) **Uncertainty in Informativeness:** Intuitively, the most informative samples would be those that lie on the decision boundaries between different classes. We first sort the distances between the embedding of the new sample to all existing class centroids in ascending order: d_n^{sorted}. Then the informativeness of a new sample is defined as the ratio between the first two nearest distances:

$$U_{info} = d_1^{sorted} / d_2^{sorted}. \tag{8}$$

These two types of uncertainty regarding new sample and class centroids are further illustrated in Fig. 8.

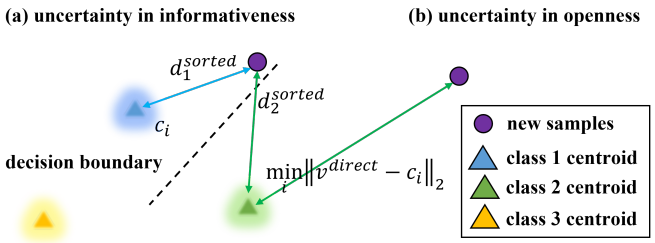


Fig. 8: **Active exploration illustration** of (a) uncertainty in informativeness, and (b) uncertainty in openness.

3.4.2 Active Sample Selection

At each time step when new data are encountered, the selection of samples for active annotations is based on the combination of both openness and informativeness uncertainty estimation using the energy-based model [10]. Similar to [59], we can express the free energy function $E(\cdot)$ of new sample v^{direct} as follows:

$$E(v^{direct}) = -T_{act} \cdot \log \sum_{i}^{K} e^{U_{open} \cdot U_{info}/T_{act}}, \quad (9)$$

where T_{act} is the temperature for controlling the trade-off between precision and recall in active selection.

3.4.3 Human-in-the-Loop Annotation

After selecting these "open" yet "informative" samples, we obtain the semantic labels of these samples by querying human annotators. In the real-world applications [89], it can be executed through online crowdsourcing platform with quality control. Since different human annotators have different preferences for naming unknown categories, it is crucial to maintain consistency between different samples of the same unknown category.

3.4.4 Model Update

Assume that we have an existing classifier $\phi_t(\cdot)$ at time step t with weight vectors $\{w_i\}_{i=1}^{K}$ for K known classes. And during time step t to time step $t+1$, we will encounter Z unknown classes. Then at time step $t+1$, our new classifier $\phi_{t+1}(\cdot)$ will be a concatenation of both known class weights and unknown class weights: $\{\{w_i\}_{i=1}^{K}, \{w_j\}_{j=K+1}^{K+Z}\}$. Both weights will be updated by the obtained active human annotations.

Note that model update is not our main contribution; there are several feasible instantiations. Since we adopt cosine classifer described in Sec. 3.5, the weight of classifier and the embedding of sample can be transformed interchangeably [48], [51]. Though more sophisticated methods (*e.g.* learning another network to generate the classifier weights for unknown classes) can be applied, here the classifier weights for unknown classes are simply hallucinated through a weighted average of the meta-embeddings from the actively selected samples for that class, where the weight is determined by $E(v^{direct})$. This classifier hallucination approach is extremely suitable for off-the-shelf deployment.

3.4.5 Dynamic Recognition Loop

To accommodate for the dynamic nature of the visual world, this procedure of active sample selection, human-in-the-loop annotation, and model update repeats each time new batch of open data are encountered. Our framework maximizes learning and recognition efficiency by taking the best from both humans and machines within a synergistic collaboration, taking care of both the long-tailed and open-ended distribution existing in the natural world.

In our implementation, the feature extractor is fixed for fast adaptation, while the visual memory is updated to accommodate for the continual stream of unknown classes during the dynamic learning loop. We have further clarified in the revised paper.

3.5 Learning

Cosine Classifier. We adopt the cosine classifier [48], [51] to produce the final classification results. Specifically, we normalize the meta-embeddings $\{v_n^{meta}\}$, where n stands for the n-th input as well as the weight vectors $\{w_i\}_{i=1}^{K}$ of the classifier $\phi(\cdot)$ (no bias term):

$$\begin{aligned} v_n^{meta} &= \frac{\|v_n^{meta}\|^2}{1+\|v_n^{meta}\|^2} \cdot \frac{v_n^{meta}}{\|v_n^{meta}\|}, \\ w_k &= \frac{w_k}{\|w_k\|}. \end{aligned} \quad (10)$$

The normalization strategy for the meta-embedding is a non-linear squashing function [90] which ensures that vectors of small magnitude are shrunk to almost zeros while vectors of big magnitude are normalized to the length slightly below 1. This function helps amplify the effect of the reachability γ (cf. Eq. (3)).

Loss Function. Since all our modules are differentiable, our model can be trained end-to-end by alternatively updating the centroids $\{c_i\}_{i=1}^{K}$ and the *dynamic meta-embedding* v_n^{meta}. The final loss function L is a combination of the cross-entropy classification loss L_{CE} and the large-margin loss between the embeddings and the centroids L_{LM}:

$$L = \sum_{n=1}^{N} L_{CE}(v_n^{meta}, y_n) + \lambda \cdot L_{LM}(v_n^{meta}, \{c_i\}_{i=1}^{K}), \quad (11)$$

where λ is set to 0.1 in our experiments via observing the accuracy curve on validation set.

Specifically, L_{CE} is the cross-entropy loss between dynamic meta-embedding v_n^{meta} and the ground truth category label y_n:

$$\begin{aligned} L_{CE}(v_n^{meta}, y_n) = y_n \log(\phi(v_n^{meta})) \\ + (1-y_n)\log(1-\phi(v_n^{meta})), \end{aligned} \quad (12)$$

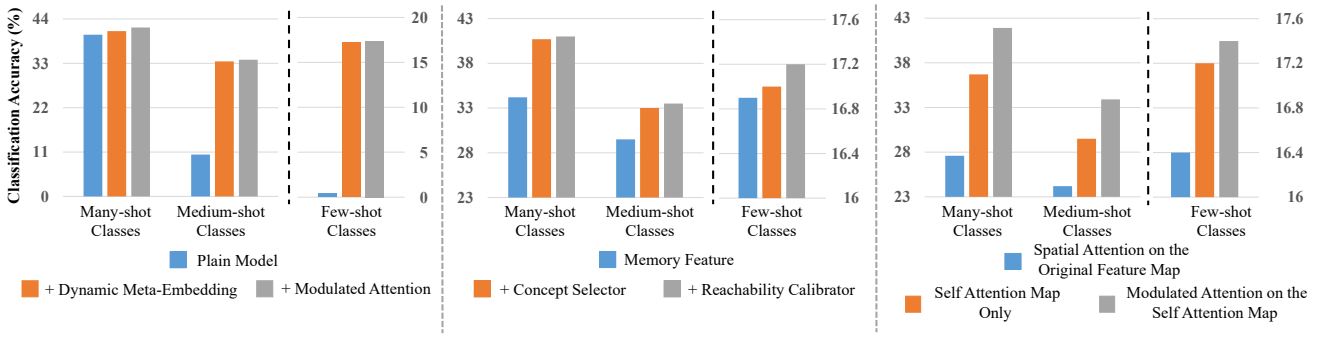
where $\phi(\cdot)$ is the cosine classifier described in Eqn. 6 in the main paper. Next we introduce the large margin loss L_{LM} between the embedding v_n^{meta} and the centroids $\{c_i\}_{i=1}^{K}$:

$$\begin{aligned} L_{LM}(v_n^{meta}, \{c_i\}_{i=1}^{K}) = \max(0, \sum_{i=y_n} \|v_n^{meta} - c_i\| \\ - \sum_{i \neq y_n} \|v_n^{meta} - c_i\| + m), \end{aligned} \quad (13)$$

where m is the margin and we set it as 5.0 in our experiments. With this formulation, we minimize the distance between each embedding and the centroid of its group and meanwhile maximize the distance between the embedding and the centroids it does not belong to.

4 Experiments

Datasets. We curated three open long-tailed benchmarks, ImageNet-LT (object-centric), Places-LT (scene-centric), and MS1M-LT (face-centric), respectively.



(a) Ablation Study of the Overall Framework (b) Ablation Study of Dynamic Meta-Embedding (c) Ablation Study of Modulated Attention

Fig. 9: **Results of ablation study.** Dynamic meta-embedding contributes most on medium-shot and few-shot classes while modulated attention helps maintain the discrimination of many-shot classes. The performance is reported with *open-set* top-1 classification accuracy on ImageNet-LT.

1) ImageNet-LT: We constructed a long-tailed version of the original ImageNet-2012 [2] by sampling a subset following Pareto distribution with power value α=6. Overall, it has 115.8K images from 1000 categories, with maximally 1280 images per class and minimally 5 images per class. The additional classes of images in ImageNet-2010 are tagged as the open set.
2) Places-LT: It contains 184.5K images from 365 categories, with the maximum of 4980 images per class and the minimum of 5 images per class. The gap between the head and tail classes are even larger than ImageNet-LT. The test images from Places-Extra69 are tagged as the additional open-set.
3) MS1M-LT: To create a long-tailed version of the MS1M-ArcFace dataset [13], [92], we sampled images for each identity with a probability proportional to the image numbers of each identity. It has 887.5K images and 74.5K identities, with a long-tailed distribution. To inspect the generalization ability of our approach, the performance is evaluated on the MegaFace benchmark [93], which has no identity overlap with MS1M-ArcFace.

Network Architectures. Following [46], [47], [48], we employed a ResNet-10 [15] trained from scratch as our backbone network for ImageNet-LT. To make a fair comparison with [6], the pre-trained ResNet-152 [15] was used as the backbone network for Places-LT. For MS1M-LT, we used the popular pre-trained ResNet-50 [15] as the backbone network. To compare with the recent works, we also adopted the two-headed ResNet-50 backbone following [22].

Evaluation Metrics. The evaluation is on the performance of each method under both the *closed-set* (test set contains no unknown classes) and *open-set* (test set contains unknown classes) settings. Under each setting, besides the overall top-1 classification accuracy [48] over all classes, we also calculated the accuracy of three disjoint subsets: *many-shot classes* (classes each with over training 100 samples), *medium-shot classes* (classes each with 20∼100 training samples) and *few-shot classes* (classes under 20 training samples). For the *open-set* setting, the *F-measure* was also reported for a balanced treatment of precision and recall following [56]. For determining open classes, the $softmax$ probability threshold was initially set as 0.1, while a more detailed analysis is provided in Sec. 4.6.

Comparison Methods. We chose for comparison state-of-the-art methods from different fields dealing with the open long-tailed data, including: (1) *metric learning*: Lifted Loss [33], (2) *hard negative mining*: Focal Loss [16], (3) *feature regularization*: Range Loss [36], (4) *few-shot learning*: FSLwF [48], (5) *long-tailed modeling*: MetaModelNet [6], and (6) *open-set detection*: Open Max [56]. We applied these methods on the same backbone networks as ours for a fair comparison. We also enabled them with class-aware mini-batch sampling [95] for effective learning. Since Model Regression [96] and MetaModelNet [6] are the most related to our work, we recorded our results along with the numbers reported in the original papers. We also included the recent advances (*e.g.* CB Focal [17], LDAM [19], Decoupling [20], BBN [21], and RIDE [22]) in long-tailed recognition for a comprehensive evaluation.

4.1 Ablation Study

Effectiveness of the Dynamic Meta-Embedding. From Fig. 9 (b), we observe that the combination of the memory feature and concept selector led to large improvements on all three shots, because the obtained memory feature transferred useful visual concepts among classes. Another observation is that the confidence calibrator is the most effective on few-shot classes. And the reachability estimation inside the confidence calibrator helped distinguish tail classes from open classes.

Effectiveness of the Modulated Attention. We observe from Fig. 9 (a) that, compared to medium-shot classes, the modulated attention contributed more to the discrimination between many-shot and few-shot classes in the experiments. Fig. 9 (c) further validates that the modulated attention is more effective than directly applying spatial attention on feature maps. It implies that adaptive contexts selection is easier to learn than the conventional feature selection.

Effectiveness of the Reachability Calibration. To further demonstrate the merit of reachability calibration for open-world setting, we conducted additional experiments following the standard settings in [58], [91] (CIFAR100 + TinyImageNet(resized)). Our approach shows favorable performance over standard open-set methods [58], [91], as listed in Table 5.

4.2 OLTR++ Benchmarking Results

In this section, we extensively evaluate the performance of various representative methods on our OLTR++ benchmarks.

ImageNet-LT. Table 4 (a) shows the performance comparison of different methods. We have the following observations. Firstly,

Backbone Net	closed-set setting				open-set setting			
ResNet-10	$\geqslant 100$	< 100 & > 20	$\leqslant 20$		$\geqslant 100$	< 100 & > 20	$\leqslant 20$	
Methods	**Many-shot**	**Medium-shot**	**Few-shot**	**Overall**	**Many-shot**	**Medium-shot**	**Few-shot**	**F-measure**
Plain Model [15]	40.9	10.7	0.40	20.9	40.1	10.4	0.40	0.295
Lifted Loss [33]	35.8	30.4	17.9	30.8	34.8	29.3	17.4	0.374
Focal Loss [16]	36.4	29.9	16.0	30.5	35.7	29.3	15.6	0.371
Range Loss [36]	35.8	30.3	17.6	30.7	34.7	29.4	17.2	0.373
+ OpenMax [56]	-	-	-	-	35.8	30.3	17.6	0.368
FSLwF [48]	40.9	22.1	15.0	28.4	40.8	21.7	14.5	0.347
CB Focal [17]	35.0	27.9	21.4	29.7	34.3	27.4	**21.1**	0.361
LDAM [19]	47.1	31.7	20.9	36.2	**46.8**	31.4	20.6	0.424
Decoupling [20]	-	-	-	41.4	-	-	-	-
Ours	47.8	38.4	18.4	39.6	44.2	**35.2**	17.5	**0.446**
ResNet-50								
Plain Model [15]	64.0	33.8	5.8	41.6	-	-	-	-
Decoupling [20]	57.1	45.2	29.3	47.7	-	-	-	-
RIDE [22]	**65.8**	51.0	34.6	54.4	-	-	-	-
Ours†	65.5	**51.8**	**35.2**	**55.0**	-	-	-	-

(a) Top-1 classification accuracy on ImageNet-LT.

Backbone Net	closed-set setting				open-set setting			
ResNet-152	$\geqslant 100$	< 100 & > 20	$\leqslant 20$		$\geqslant 100$	< 100 & > 20	$\leqslant 20$	
Methods	**Many-shot**	**Medium-shot**	**Few-shot**	**Overall**	**Many-shot**	**Medium-shot**	**Few-shot**	**F-measure**
Plain Model [15]	**45.9**	22.4	0.36	27.2	**45.9**	22.4	0.36	0.366
Lifted Loss [33]	41.1	35.4	24.0	35.2	41.0	35.2	23.8	0.459
Focal Loss [16]	41.1	34.8	22.4	34.6	41.0	34.8	22.3	0.453
Range Loss [36]	41.1	35.4	23.2	35.1	41.0	35.3	23.1	0.457
+ OpenMax [56]	-	-	-	-	41.1	35.4	23.2	0.458
FSLwF [48]	43.9	29.9	29.5	34.9	38.1	19.5	14.8	0.375
CB Focal [17]	43.4	39.1	**30.5**	38.9	42.3	37.7	**28.8**	0.490
LDAM [19]	45.6	37.8	23.9	35.7	45.6	37.7	23.5	0.485
Decoupling [20]	40.6	39.1	28.6	37.3	-	-	-	-
Ours	44.0	**40.6**	28.5	**39.3**	43.7	**40.2**	28.0	**0.500**

(b) Top-1 classification accuracy on Places-LT.

TABLE 4: **Benchmarking results on (a) ImageNet-LT and (b) Places-LT.** Our approach provides a comprehensive treatment to all the many/medium/few-shot classes as well as the open classes with substantial advantages on all aspects. † denotes using the two-headed ResNet-50 backbone following [22].

Method	FPR (95% TPR)	Detection Error
Softmax Pred. [91]	82.2	43.6
Ours	51.5	29.9
ODIN [58]†	44.3	24.6
Energy OOD [59]†	40.7	21.1
Ours†	**35.4**	**18.0**

TABLE 5: **Open class detection error (%) comparison.** It is performed on the standard open-set benchmark, CIFAR100 + TinyImageNet (resized). "†" denotes the setting where open samples are used to tune algorithmic parameters.

both Lifted Loss [33] and Focal Loss [16] greatly boosted the performance of few-shot classes by enforcing feature regularization. However, they also sacrificed the performance on many-shot classes since there are no built-in mechanism of adaptively handling samples of different shots. Secondly, OpenMax [56] improved the results under the open-set setting. However, the accuracy degraded when it was evaluated with *F-measure*, which considers both precision and recall in open-set. This proves that when the open classes are compounded with the tail classes, it becomes challenging to perform the distribution fitting that [56] requires. Lastly, although the few-shot learning without forgetting approach [48] retained the many-shot class accuracy, it had difficulty dealing with the imbalanced base classes which are lacked in the current few-shot paradigm. Overall, as demonstrated in Fig. 10, our approach provides a comprehensive treatment to all the many/medium/few-shot classes as well as the open classes with substantial improvements on all aspects.

Places-LT. Similar observations can be made on the Places-LT benchmark as shown in Table 4 (b). With a much stronger baseline (*i.e.* ImageNet pre-trained ResNet-152), our approach still consistently outperformed other alternatives under both the closed-set and open-set settings. The advantage is even more profound under *F-measure*.

MS1M-LT. We trained on the MS1M-LT dataset and report results on the MegaFace identification track, which is a standard benchmark in the face recognition field. Since the face identities in the training set and the test set are disjoint, we adopted an indirect way to partition the testing set into subsets of different shots. We approximated the pseudo shots of each test sample by counting the number of training samples that are similar to it by at least a threshold (feature similarity greater than 0.7). Apart from many-shot, few-shot, one-shot subsets, we also obtained a zero-shot subset, for which we could not find any sufficiently similar samples in the training set. It can be observed that our approach has the most advantage on one-shot identities (3.0% gains) and zero-shot identities (1.8% gains) as shown in Table 6.

SUN-LT. To directly compare with [96] and [6], we also tested out method on the SUN-LT benchmark they provided. The final results are listed in Table 8. Instead of learning a series of classifier transformations, our approach transfered visual knowledge among features and achieved an 1.4% improvement over the prior best.

Backbone Net ResNet-50 **Methods**	**MegaFace Identification Rate** $\geqslant 5$ **Many**	< 5 & $\geqslant 2$ **Few**	< 2 & $\geqslant 1$ **One-shot**	$= 0$ **Zero-shot**	**Full Test**	Gender Sub-Groups **Male**	**Female**	Ethnicity Sub-Groups **Caucasian**	**Asian**	**Indian**	**African**
Plain Model [15]	80.64	71.98	84.60	77.72	73.88	78.30	78.70	85.83	75.67	76.42	79.28
Range Loss [36]	78.60	71.36	83.14	77.40	72.17	78.12	77.45	86.11	74.86	75.94	76.37
Fair Feature [79]	-	-	-	-	-	78.23	77.61	86.34	74.97	76.25	77.62
Debiasing [94]	-	-	-	-	-	78.73	78.85	86.36	75.89	76.90	79.77
Ours	**80.82**	**72.44**	**87.60**	**79.50**	**74.51**	**79.04**	**79.08**	**86.59**	**76.22**	**77.05**	**80.37**

TABLE 6: **Benchmarking results on MegaFace.** Our approach achieved the best performance on natural-world datasets when compared to other state-of-the-art methods. Furthermore, our approach had across-board improvements on 'male' and 'female' gender sub-groups as well as 'Caucasian','Asian', 'Indian' and 'African' ethnicity sub-groups.

Methods	CIFAR-10-LT (im. ratio = 100) **Many**	**Medium**	**Overall**	CIFAR-100-LT (im. ratio = 100) **Many**	**Medium**	**Few**	**Overall**	iNaturalist-18 **Many**	**Medium**	**Few**	**Overall**
Plain Model [15]	-	-	70.4	66.1	37.3	10.6	39.1	**72.2**	63.0	57.2	61.7
CB Focal [17]	-	-	74.6	-	-	-	36.0	-	-	-	61.1
LDAM [19]	80.5	67.0	77.0	61.5	41.7	20.2	42.0	-	-	-	64.6
Decoupling [20]	-	-	-	65.7	43.6	17.3	43.2	65.0	66.3	65.5	65.9
BBN [21]	-	-	79.8	-	-	-	42.6	49.4	70.8	65.3	66.3
RIDE [22]	-	-	-	67.9	48.4	21.8	47.0	70.2	71.3	71.7	71.4
Ours	79.7	62.0	76.3	61.8	41.4	17.6	41.2	62.4	66.7	65.9	65.3
Ours†	**84.1**	**68.3**	**80.8**	**68.0**	**48.9**	**22.6**	**47.4**	70.5	**72.0**	**72.2**	**71.9**

TABLE 7: **Benchmarking results on CIFAR-10-LT, CIFAR-100-LT and iNaturalist-18.** † denotes using the two-headed ResNet-50 backbone following [22].

Backbone Net ResNet-152 **Method**	**closed-set setting** **Top-1 Accuracy**	**Top-5 Accuracy**
Plain Model [15]	48.0	77.8
Cost-Sensitive [32]	52.4	81.9
Model Reg. [96]	54.7	82.4
MetaModelNet [6]	57.3	83.6
Ours	**58.7**	**84.2**

TABLE 8: **Benchmarking results on SUN-LT.**

	Stage 1 **Known / Unknown**	2 **Known / Unknown**
Plain Model [15]	21.4 / 10.8	18.2 / 6.6
LwF [76]	26.5 / 17.2	24.1 / 11.3
ACL [77]	30.0 / 21.2	27.9 / 16.5
Ours	**38.5 / 26.7**	**36.3 / 24.9**

TABLE 9: **Performance of dynamic recognition loop.**

Note that our approach also incured much less computational cost since MetaModelNet [6] requires a recursive training procedure.

CIFAR-10-LT & CIFAR-100-LT & iNaturalist-18. We also compared OLTR++ with recent proposed methods for long-tailed recognition (*e.g.* CB Focal [17], LDAM [19], Decoupling [20], BBN [21], and RIDE [22]) on CIFAR-10-LT [17], CIFAR-100-LT [17] and iNaturalist-18 [14]. When combining our framework with the two-headed ResNet-50 [22], it achieved state-of-the-art performance on all benchmarks, with consistent gains over many-, medium-, and few-shot classes under various imbalance ratios.

4.3 Performance on Fairness Analysis

Results. On the sensitive attribute performance on MS1M-LT, the last six columns in Table 6 show that our approach achieved across-board improvements on both gender sub-groups and ethnicity sub-groups, which has an encouraging implication for effective fairness learning.

Indications. The goal of evaluating the performance on different gender sub-groups and ethnicity sub-groups here is to draw connection between long-tailed recognition and the larger community of bias and fairness in artificial intelligence, which could possibly motivate more future research at the intersection of these two. To provide a comprehensive comparison on this interdisciplinary sub-field, we further implement and include two representative methods in fairness learning (*i.e.* fair feature [79] and latent debiasing [94]) into the evaluation.

4.4 Performance on Active Exploration

4.4.1 *Effectiveness of Active Sample Selection*

Setting. The experiments were performed on ImageNet-LT (as the initial labeled pool), with the additional classes of images in ImageNet-2010 as the open/exploration set. This open/exploration set was further mixed with images with known classes, which were taken from the original ImageNet dataset excluding ImageNet-LT. For fair comparisons, all the methods adopted the same backbone network and were benchmarked under different percentages of selected open samples for oracle annotations (from 10% to 30%). The evaluation metric was the average recognition accuracy over all the open classes.

Results. We evaluated the performance of OLTR++ to recognize the open classes under the active exploration setting. We compared our results with several representative methods in active learning, including random sampling, Bayesian-uncertainty-based method DBAL [68], core-set-based method CoreSet [97] and adversarial-learning-based method VAAL [69]. As shown in Fig. 12 (a), our OLTR++ validates its advantage over different percentages of selected open samples for active annotation and demonstrates dynamic-world learning efficiency.

Analysis. To further understand the active exploration ability of OLTR++, we visualized the samples of selected and not selected images by our active exploration approach in Fig. 12 (b). We can observe that OLTR++ tends to select canonical images with representative parts and appearance for unseen classes.

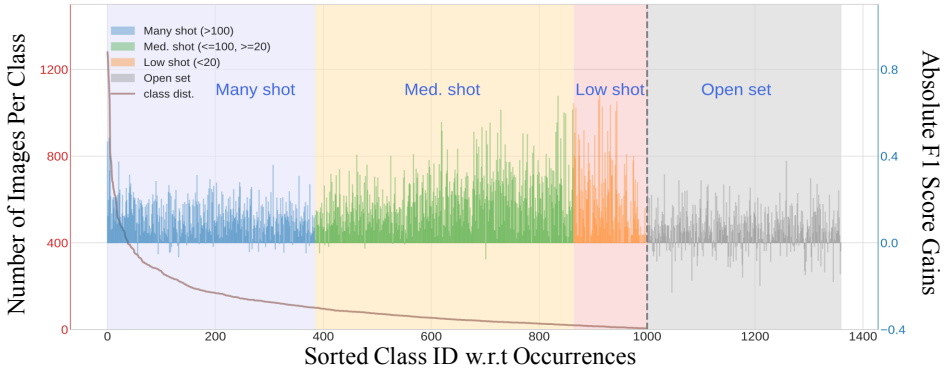


Fig. 10: **The absolute F1 score of our method over the plain model.** Ours has across-the-board performance gains w.r.t. many/medium/few-shot and open classes.

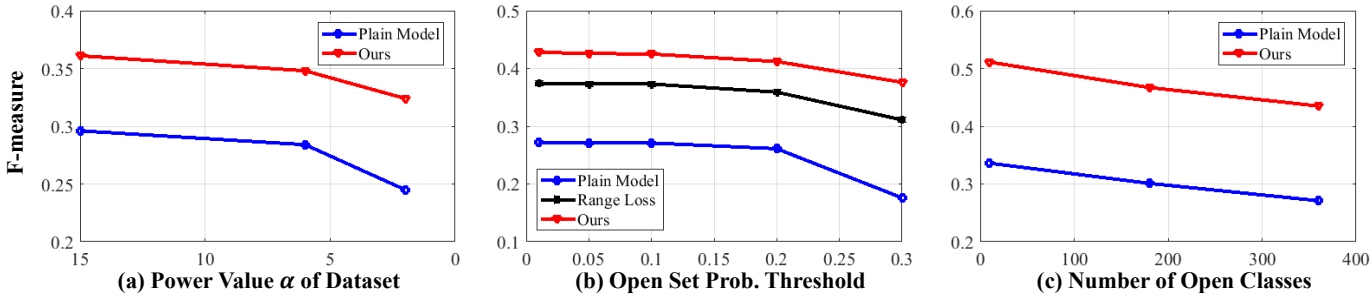


Fig. 11: **The influence of (a) dataset longtail-ness, (b) open-set probability threshold, and (c) the number of open classes.** As the dataset became more imbalanced, our approach only underwent a moderate performance drop. Our approach also demonstrates great robustness to the contamination of open classes.

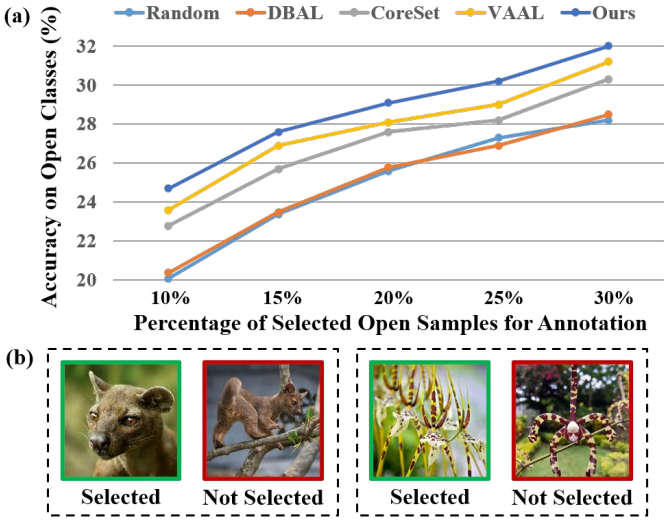


Fig. 12: (a) **Performance of active exploration on open classes.** Our OLTR++ validates its advantage over different percentages of selected open samples for active annotation. (b) **Samples of selected images** by our active exploration approach.

4.4.2 *Effectiveness of Dynamic Recognition Loop*

Setting. The experiments were performed in a similar setting to active exploration, except that the open/exploration set was divided into two subsets to mimic the two stages in incremental/continual learning scenario. Specifically, both Stage 1 and Stage 2 have splits from the additional classes of images in ImageNet-2010 as the open/exploration set. For a fair comparison, the base data is not used for model update in the incremental step. To mimic the real-world applications, the accuracies of "known/unknown" is calculated by classifying test samples to a combination of known and unknown labels.

Results. We compared our results with several representative methods in incremental/continual learning, including Learning without Forgetting (LwF) [76], and Adversarial Continual Learning (ACL) [77]. LwF [76] is a seminal work in the field of incremental learning, which serves as a strong and must-have baseline by employing a soft multi-task learning paradigm. ACL [77] is a recent adversarial-learning-based method of state-of-the-art performance without using external data or models. It is also equipped with open-sourced code and evaluation scripts for fair comparisons. As listed in Table 9, our OLTR++ maintains the learning and recognition effectiveness during the dynamic loop.

4.5 Influence of Tailness and Openness

Influence of Dataset Longtail-ness. The longtail-ness of the dataset (*e.g.* the degree of imbalance of the class distribution) can have a negative impact on the model performance. For faster investigating, the weights of the backbone network were frozen during training. From Fig. 11 (a), we observe that as the dataset became more imbalanced (*i.e.* power value α decreases), our approach only underwent a moderate performance drop. In other words, dynamic meta-embedding enables effective knowledge transfer among data-abundant and data-scarce classes.

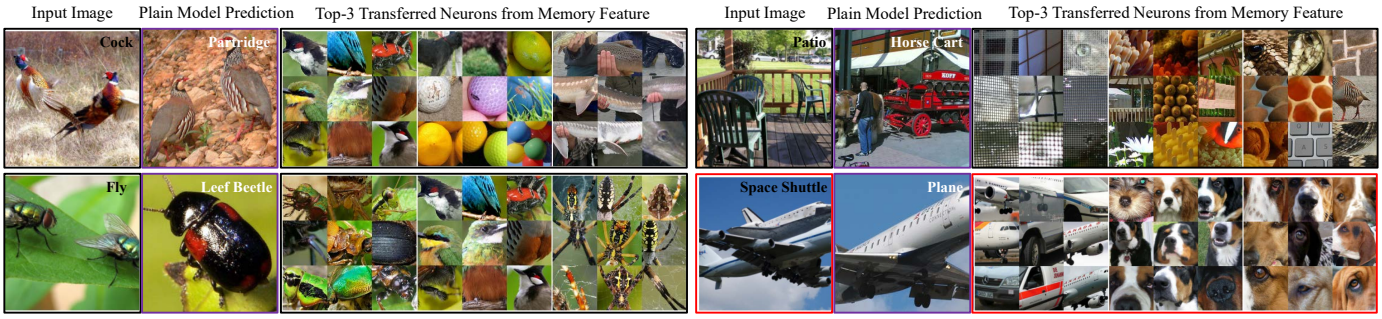


Fig. 13: **Examples of the top-3 infused visual concepts from memory feature.** Except for the bottom right failure case (marked in red), all the other three input images were misclassified by the plain model and correctly classified by our model. For example, to classify the top left image which belongs to a tail class 'cock', our approach learned to transfer visual concepts that represents "bird head", "round shape" and "dotted texture" respectively.

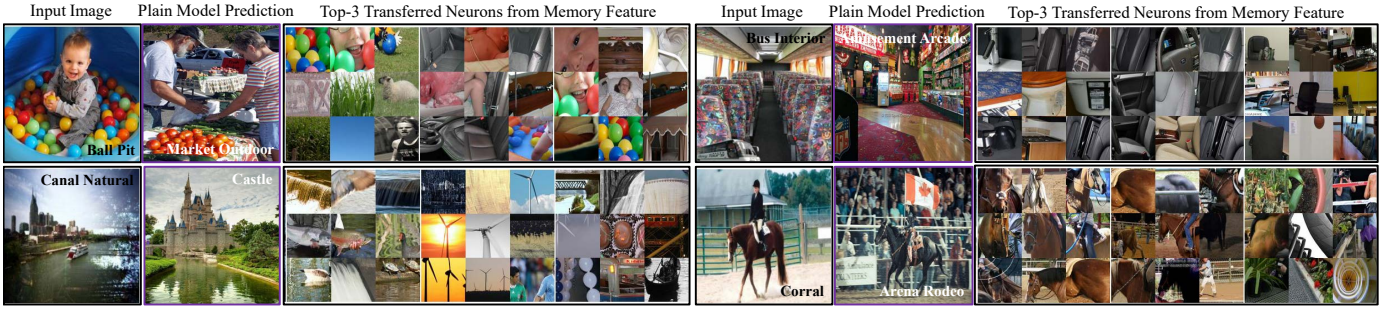


Fig. 14: **Examples of the infused visual concepts from "memory feature" in Places-LT.**

Method	Top-1 Accuracy	Top-5 Accuracy
Random	57.9	63.2
ImageNet Pre-Train [2]	71.3	80.6
Ours	**77.4**	**85.8**

TABLE 10: **Performance comparison of unsupervised attribute discovery.** It is evaluated on the CelebA dataset [26].

Influence of Open-Set Probabilistic Threshold. The performance change w.r.t. the open-set probability threshold is demonstrated in Fig. 11 (b). Compared to the plain model [15] and range loss [36], the performance of our approach changed steadily as the open-set threshold rose. The reachability estimator in our framework helped calibrate the sample confidence, thus enhanced robustness to open classes.

Influence of the Number of Open Classes. Finally we investigate performance change w.r.t. the number of open classes. Fig. 11 (c) indicates that our approach demonstrates great robustness to the contamination of open classes.

4.6 Further Analysis

In this section, we visualize and interpret memory feature in our framework, as well as present the relation of OLTR++ to fairness analysis and typical failure cases.

Infused Memory Feature. In Fig. 13, we inspect the visual concepts that memory feature has infused by visualizing its top activating neurons. Specifically, for each input image, we extracted its top-3 transferred neurons in memory feature. And each neuron is visualized by a collection of highest activated patches [98] over the whole training set. For example, to classify the top left image which belongs to a tail class 'cock', our approach learned to transfer visual concepts that represents "bird head", "round shape" and "dotted texture" respectively. After feature infusion, the dynamic meta-embedding became more informative and discriminative.

We visualize the "memory feature" in Places-LT similarly to ImageNet-LT. Examples of the infused visual concepts from "memory feature" in Places-LT are presented in Fig. 14. We observe that "memory feature" indeed encodes discriminative visual traits for the underlying scene.

Following [26], we visualize the "memory feature" in MS1M-LT by contrasting the least activated average image and the most activated average image of the top firing neuron. From Fig. 15, we observe that "memory feature" in MS1M-LT infused several identity-related attributes (*e.g.* "high cheekbones", "dark skin color" and "narrow eyes") for precise recognition.

Unsupervised Attribute Discovery. We also quantitatively evaluate the "memory feature" in MS1M-LT by performing unsupervised attribute discovery experiment. The performance is reported on the CelebA dataset [26] via linear probing, as listed in Table 10. Our approach outperforms randomly initialized and ImageNet pre-trained features on both top-1 and top-5 average accuracy.

Failure Cases. Since our approach encourages the feature infusion among classes, it slightly sacrifices the fine-grained discrimination for the promotion of under-representative classes. One typical failure case of our approach is the confusion between many-shot and medium-shot classes. For example, the bottom right image in Fig. 13 was misclassified into 'airplane' because some cross-category traits like "nose shape" and "eye shape" were infused. Feature disentanglement [99] and strong contrastive learning [100] are potential alleviations to this trade-off issue.

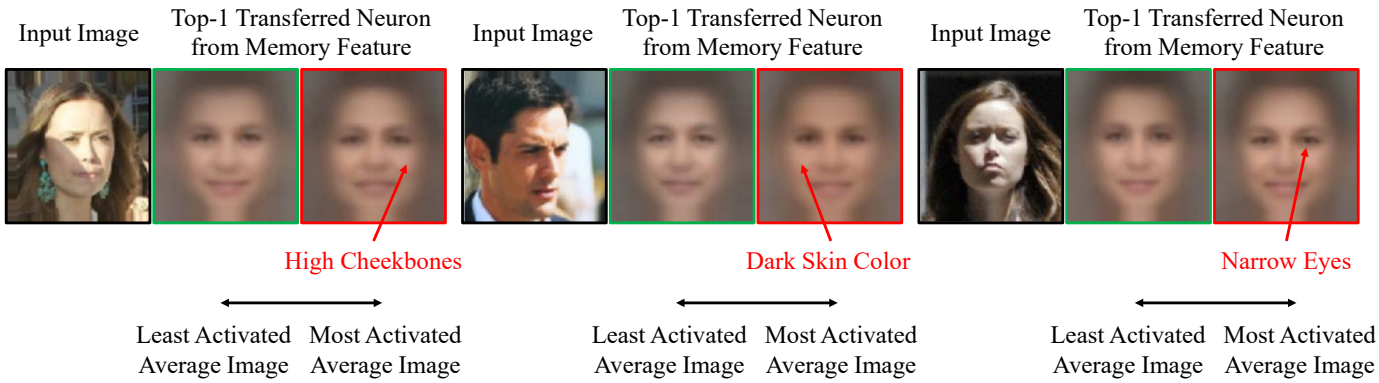


Fig. 15: **Examples of the infused visual concepts from "memory feature" in MS1M-LT.**

5 Conclusions

We introduce the OLTR++ task that learns from natural long-tail open-end distributed data and optimizes the overall accuracy over a balanced test set. We propose an integrated OLTR++ algorithm, dynamic meta-embedding, in order to share visual knowledge between head and tail classes and to reduce confusion between tail and open classes. We validated our method on three curated large-scale OLTR++ benchmarks (ImageNet-LT, Places-LT and MS1M-LT) as well as active exploration. Our work can enable future researches that are directly transferable to real-world applications.

References

[1] W. J. Reed, "The pareto, zipf and other power laws," *Economics letters*, 2001. 1

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009. 1, 2, 5, 8, 12

[3] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014. 1

[4] O. Vinyals, C. Blundell, T. Lillicrap, and D. Wierstra, "Matching networks for one shot learning," in *NeurIPS*, 2016. 1, 3

[5] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, 2015. 1

[6] Y.-X. Wang, D. Ramanan, and M. Hebert, "Learning to model the tail," in *NeurIPS*, 2017. 1, 2, 8, 9, 10

[7] Z. Miao, K. M. Gaynor, J. Wang, Z. Liu, O. Muellerklein, M. S. Norouzzadeh, A. McInturff, R. C. Bowie, R. Nathan, X. Y. Stella *et al.*, "Insights and approaches using deep learning to classify wildlife," *Nature - Scientific reports*, 2019. 1

[8] Y. Duan, J. Schulman, X. Chen, P. L. Bartlett, I. Sutskever, and P. Abbeel, "Rl^2: Fast reinforcement learning via slow reinforcement learning," *arXiv preprint arXiv:1611.02779*, 2016. 1, 3

[9] J. Ba, G. E. Hinton, V. Mnih, J. Z. Leibo, and C. Ionescu, "Using fast weights to attend to the recent past," in *NeurIPS*, 2016. 1, 3

[10] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang, "A tutorial on energy-based learning," *Predicting structured data*, vol. 1, no. 0, 2006. 2, 7

[11] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," *arXiv preprint arXiv:1711.07971*, 2017. 2, 6

[12] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *TPAMI*, 2018. 2

[13] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *ECCV*, 2016. 2, 8

[14] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie, "The inaturalist species classification and detection dataset," in *CVPR*, 2018. 2, 10

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016. 3, 8, 9, 10, 12

[16] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *ICCV*, 2017. 2, 3, 8, 9

[17] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *CVPR*, 2019. 2, 3, 8, 9, 10

[18] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, "Large-scale long-tailed recognition in an open world," in *CVPR*, 2019. 2, 3

[19] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *NeurIPS*, 2019. 2, 3, 8, 9, 10

[20] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis, "Decoupling representation and classifier for long-tailed recognition," in *ICLR*, 2020. 2, 3, 8, 9, 10

[21] B. Zhou, Q. Cui, X.-S. Wei, and Z.-M. Chen, "Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition," in *CVPR*, 2020. 3, 8, 10

[22] X. Wang, L. Lian, Z. Miao, Z. Liu, and S. X. Yu, "Long-tailed recognition by routing diverse distribution-aware experts," in *ICLR*, 2021. 3, 8, 9, 10

[23] R. Salakhutdinov, A. Torralba, and J. Tenenbaum, "Learning to share visual appearance for multiclass object detection," in *CVPR*, 2011. 2

[24] X. Zhu, D. Anguelov, and D. Ramanan, "Capturing long-tail distributions of object subcategories," in *CVPR*, 2014. 2

[25] S. Bengio, "The battle against the long tail," in *Talk on Workshop on Big Data and Statistical Machine Learning*, 2015. 2

[26] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *ICCV*, 2015. 2, 12

[27] X. Zhu, C. Vondrick, C. C. Fowlkes, and D. Ramanan, "Do we need more training data?" *IJCV*, 2016. 2

[28] W. Ouyang, X. Wang, C. Zhang, and X. Yang, "Factors in finetuning deep model for object detection with long-tail distribution," in *CVPR*, 2016. 2

[29] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *CVPR*, 2016. 2

[30] G. Van Horn and P. Perona, "The devil is in the tails: Fine-grained classification in the wild," *arXiv preprint arXiv:1709.01450*, 2017. 2

[31] Y. Cui, Y. Song, C. Sun, A. Howard, and S. Belongie, "Large scale fine-grained categorization and domain-specific transfer learning," in *CVPR*, 2018. 2

[32] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Learning deep representation for imbalanced classification," in *CVPR*, 2016. 2, 10

[33] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *CVPR*, 2016. 2, 8, 9

[34] Q. Dong, S. Gong, and X. Zhu, "Class rectification hard mining for imbalanced deep learning," in *ICCV*, 2017. 2

[35] D. Ha, A. Dai, and Q. V. Le, "Hypernetworks," *arXiv preprint arXiv:1609.09106*, 2016. 2

[36] X. Zhang, Z. Fang, Y. Wen, Z. Li, and Y. Qiao, "Range loss for deep face recognition with long-tailed training data," in *CVPR*, 2017. 2, 8, 9, 10, 12

[37] H.-J. Ye, H.-Y. Chen, D.-C. Zhan, and W.-L. Chao, "Identifying and compensating for feature deviation in imbalanced deep learning," *arXiv preprint arXiv:2001.01385*, 2020. 2

[38] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar, "Long-tail learning via logit adjustment," *arXiv preprint arXiv:2007.07314*, 2020. 2

[39] J. Schmidhuber, "A neural network that embeds its own meta-levels," in *ICNN*, 1993. 3

[40] L. Bertinetto, J. F. Henriques, J. Valmadre, P. Torr, and A. Vedaldi, "Learning feed-forward one-shot learners," in *NeurIPS*, 2016. 3

[41] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *ICLR*, 2017. 3

[42] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *ICML*, 2016. 3, 5

[43] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," *arXiv preprint arXiv:1703.03400*, 2017. 3

[44] F. S. Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *CVPR*, 2018. 3

[45] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *NeurIPS*, 2017. 3, 5

[46] B. Hariharan and R. B. Girshick, "Low-shot visual recognition by shrinking and hallucinating features." in *ICCV*, 2017. 3, 8

[47] Y.-X. Wang, R. Girshick, M. Hebert, and B. Hariharan, "Low-shot learning from imaginary data," *arXiv preprint arXiv:1801.05401*, 2018. 3, 8

[48] S. Gidaris and N. Komodakis, "Dynamic few-shot visual learning without forgetting," in *CVPR*, 2018. 3, 7, 8, 9

[49] M. Ren, R. Liao, E. Fetaya, and R. S. Zemel, "Incremental few-shot learning with attention attractor networks," *arXiv preprint arXiv:1810.07218*, 2018. 3

[50] S. Qiao, C. Liu, W. Shen, and A. Yuille, "Few-shot image recognition by predicting parameters from activations," in *CVPR*, 2018. 3

[51] H. Qi, M. Brown, and D. G. Lowe, "Low-shot learning with imprinted weights," in *CVPR*, 2018. 3, 7

[52] G. E. Hinton and D. C. Plaut, "Using fast weights to deblur old memories," in *Proceedings of the ninth annual conference of the Cognitive Science Society*, 1987. 3

[53] J. Schmidhuber, "Learning to control fast-weight memories: An alternative to dynamic recurrent networks," *Neural Computation*, 1992. 3

[54] T. Munkhdalai and H. Yu, "Meta networks," *arXiv preprint arXiv:1703.00837*, 2017. 3, 5

[55] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," *TPAMI*, 2013. 3

[56] A. Bendale and T. E. Boult, "Towards open set deep networks," in *CVPR*, 2016. 3, 8, 9

[57] T. DeVries and G. W. Taylor, "Learning confidence for out-of-distribution detection in neural networks," *arXiv preprint arXiv:1802.04865*, 2018. 3

[58] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," in *ICLR*, 2018. 3, 8, 9

[59] W. Liu, X. Wang, J. D. Owens, and Y. Li, "Energy-based out-of-distribution detection," in *NeurIPS*, 2020. 3, 7, 9

[60] A. Bendale and T. Boult, "Towards open world recognition," in *CVPR*, 2015. 3

[61] T. E. Boult, S. Cruz, A. R. Dhamija, M. Gunther, J. Henrydoss, and W. J. Scheirer, "Learning and the unknown: Surveying steps toward open world recognition," in *AAAI*, 2019. 3

[62] K. Joseph, S. Khan, F. S. Khan, and V. N. Balasubramanian, "Towards open world object detection," in *CVPR*, 2021. 3

[63] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *TPAMI*, 2013. 3

[64] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly," *TPAMI*, 2018. 3

[65] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha, "Synthesized classifiers for zero-shot learning," in *CVPR*, 2016. 3

[66] J.-J. Zhu and J. Bento, "Generative adversarial active learning," *arXiv preprint arXiv:1702.07956*, 2017. 3

[67] C. Mayer and R. Timofte, "Adversarial sampling for active learning," in *WACV*, 2020. 3

[68] Y. Gal, R. Islam, and Z. Ghahramani, "Deep bayesian active learning with image data," *arXiv preprint arXiv:1703.02910*, 2017. 3, 10

[69] S. Sinha, S. Ebrahimi, and T. Darrell, "Variational adversarial active learning," in *ICCV*, 2019. 3, 10

[70] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell, "Progressive neural networks," *arXiv preprint arXiv:1606.04671*, 2016. 3

[71] S. Lee, J. Ha, D. Zhang, and G. Kim, "A neural dirichlet process mixture model for task-free continual learning," in *ICLR*, 2019. 3

[72] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "icarl: Incremental classifier and representation learning," in *CVPR*, 2017. 3

[73] R. Aljundi, M. Lin, B. Goujaud, and Y. Bengio, "Gradient based sample selection for online continual learning," *NeurIPS*, 2019. 3

[74] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *PNAS*, 2017. 3

[75] F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence," in *ICML*, 2017. 3

[76] Z. Li and D. Hoiem, "Learning without forgetting," *TPAMI*, 2017. 3, 10, 11

[77] S. Ebrahimi, F. Meier, R. Calandra, T. Darrell, and M. Rohrbach, "Adversarial continual learning," in *ECCV*, 2020. 3, 10, 11

[78] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *The 3rd innovations in theoretical computer science conference*, 2012. 4

[79] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *ICML*, 2013. 4, 10

[80] D. Madras, E. Creager, T. Pitassi, and R. Zemel, "Learning adversarially fair and transferable representations," *arXiv preprint arXiv:1802.06309*, 2018. 4

[81] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," *arXiv preprint arXiv:1810.03993*, 2018. 4

[82] L. Anne Hendricks, K. Burns, K. Saenko, T. Darrell, and A. Rohrbach, "Women also snowboard: Overcoming bias in captioning models," in *ECCV*, 2018. 4

[83] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NeurIPS*, 2012. 5

[84] Y.-C. Hsu, Z. Lv, and Z. Kira, "Learning to cluster in order to transfer across domains and tasks," *arXiv preprint arXiv:1711.10125*, 2017. 5

[85] W. Deng, Q. Liao, L. Zhao, D. Guo, G. Kuang, D. Hu, and L. Liu, "Joint clustering and discriminative feature alignment for unsupervised domain adaptation," *TIP*, 2021. 5

[86] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *ECCV*, 2016. 5

[87] N. Savinov, A. Raichuk, R. Marinier, D. Vincent, M. Pollefeys, T. Lillicrap, and S. Gelly, "Episodic curiosity through reachability," *arXiv preprint arXiv:1810.02274*, 2018. 5

[88] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017. 6

[89] Z. Miao, Z. Liu, K. M. Gaynor, M. S. Palmer, S. X. Yu, and W. M. Getz, "Iterative human and automated identification of wildlife images," *Nature - Machine Intelligence*, 2021. 7

[90] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *NeurIPS*, 2017. 7

[91] D. Hendrycks and K. Gimpel, "Baseline for detecting misclassified and out-of-distribution examples in neural networks," in *ICLR*, 2017. 8, 9

[92] J. Deng, J. Guo, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," *arXiv preprint arXiv:1801.07698*, 2018. 8

[93] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale," in *CVPR*, 2016. 8

[94] V. V. Ramaswamy, S. S. Kim, and O. Russakovsky, "Fair attribute classification through latent space de-biasing," in *CVPR*, 2021. 10

[95] L. Shen, Z. Lin, and Q. Huang, "Relay backpropagation for effective learning of deep convolutional neural networks," in *ECCV*, 2016. 8

[96] Y.-X. Wang and M. Hebert, "Learning to learn: Model regression networks for easy small sample learning," in *ECCV*, 2016. 8, 9, 10

[97] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in *ICLR*, 2018. 10

[98] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV*, 2014. 12

[99] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *TPAMI*, 2013. 12

[100] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020. 12

Ziwei Liu is currently an Assistant Professor at Nanyang Technological University (NTU). Previously, he was a senior research fellow at the Chinese University of Hong Kong and a postdoctoral researcher at University of California, Berkeley. Ziwei received his PhD from the Chinese University of Hong Kong in 2017. His research revolves around computer vision/graphics, machine learning, and robotics. He has published over 60 papers on top-tier conferences and journals in relevant fields, including CVPR, ICCV, ECCV, AAAI, IROS, SIGGRAPH, T-PAMI, and TOG. He is the recipient of Microsoft Young Fellowship, Hong Kong PhD Fellowship, ICCV Young Researcher Award, and HKSTP best paper award.

Zhongqi Miao is currently a PhD candidate at International Computer Science Institute, University of California at Berkeley. His research interests include imbalanced classification and domain adaptation in computer vision.

Xiaohang Zhan is currently a Ph.D. student in Multimedia Laboratory, The Chinese University of Hong Kong. His research interests include computer vision and machine learning, particularly unsupervised learning.

Jiayun Wang is currently a Ph.D. candidate in the Vision Science program, on computational vision track, University of California at Berkeley. His research interests include 3D vision and unsupervised learning.

Boqing Gong received his Ph.D. in 2015 at the University of Southern California, where the Viterbi Fellowship partially supported his work. He is currently a research scientist at Google, Seattle. Before joining Google in 2019, he worked in Tencent AI Lab (Seattle) and was a tenure-track Assistant Professor at the University of Central Florida (UCF). He received an NSF CRII award in 2016 and an NSF BIGDATA award in 2017, both of which were the first of their kinds ever granted to UCF. His research interestes include data- and label-efficient learning (e.g., domain adaptation, few-shot, reinforcement, webly-supervised, and self-supervised learning), and visual analytics of objects, scenes, human activities, and their attributes.

Stella X. Yu received her Ph.D. from the School of Computer Science at Carnegie Mellon University, where she studied robotics at the Robotics Institute and vision science at the Center for the Neural Basis of Cognition. She continued her computer vision research as a postdoctoral fellow at University of California at Berkeley, and then studied art and vision as a Clare Booth Luce Professor at Boston College, during which she received an NSF CAREER award. She joined the International Computer Science Institute (ICSI) as a senior research scientist in 2012 and began leading the Vision Group in 2015. She became a Senior Fellow at the Berkeley Institute for Data Science (BIDS) at UC Berkeley in 2016. Dr. Yu is on the faculty of the Computer Science Department and the Vision Science Program at UC Berkeley, and she is also affiliated faculty with the Department of Computer and Information Science at the University of Pennsylvania. Her research interests include understanding visual perception from multiple perspectives: human vision, computer vision, robotic vision, and artistic vision.