OPEN-SET REPRESENTATION LEARNING THROUGH COMBINATORIAL EMBEDDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual recognition tasks are often limited to dealing with a small subset of classes simply because the labels for the remaining classes are unavailable. We are interested in identifying novel concepts in a dataset through the representation learning based on both labeled and unlabeled examples, and extending the horizon of recognition to both known and novel classes. To address this challenging task, we propose a combinatorial learning approach, which naturally clusters the examples in unseen classes using the compositional knowledge given by multiple supervised meta-classifiers on heterogeneous label spaces. The representations given by the combinatorial embedding are made more robust by consistency regularization. We also introduce a metric learning strategy to estimate pairwise pseudo-labels for improving the representations of unlabeled examples, which preserves semantic relations across known and novel classes effectively. The proposed algorithm discovers novel concepts via a joint optimization of enhancing the discrimitiveness of unseen classes as well as learning the representations of known classes generalizable to novel ones. Our extensive experiments demonstrate remarkable performance gains by the proposed approach in multiple image retrieval and novel class discovery benchmarks.

1 INTRODUCTION

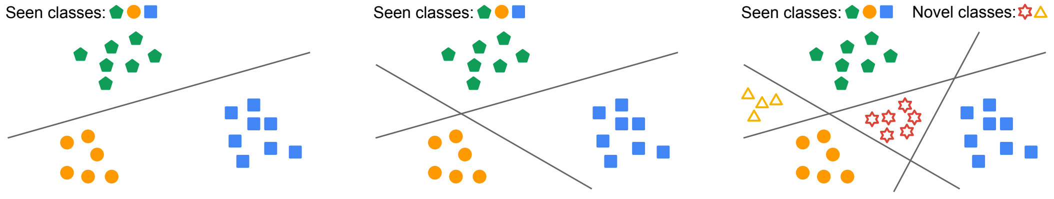
Despite the remarkable success of machine learning fueled by deep neural networks, the existing frameworks still have critical limitations in an open-world setting, where some categories are not defined a-priori and the labels for some classes are missing. Although there have been a growing number of works that identify new classes in unlabeled data given a set of labeled examples (Hsu et al., 2018; 2019; Chi et al., 2021; Brbić et al., 2020; Han et al., 2019; 2020), they often assume that all the unlabeled examples belong to unseen classes and/or the number of novel classes is known in advance, which make their problem settings unrealistic.

To address the limitation, this paper introduces an algorithm applicable to a more realistic setting. We aim to discover and learn the representations of unseen categories without any prior information or supervision about novel classes, in a natural setting that unlabeled data may contain examples in both seen and unseen classes. This task requires the model to be able to effectively identify unseen classes while preserving the information of previously seen classes.

We propose a representation learning approach based on the combinatorial classification (Seo et al., 2019), where the examples in unseen categories are identified by the composition of multiple meta-classifiers. Figure 1 illustrates the main concept of our *combinatorial embedding* framework, which forms partitions for novel classes via a combination of multiple classifiers for the meta-classes made of several known classes. The images in the same meta-class potentially have common attributes that are helpful for knowledge transfer to novel classes. The learned representations via combinatorial embedding become even stronger by consistency regularization, and more accurate representations of novel classes are obtained by transductive metric learning based on pairwise similarity.

The main contributions of our work are summarized as follows.

- We propose a novel combinatorial learning framework, which embeds the examples in both seen and novel classes effectively by the composition of the knowledge learned from multiple heterogeneous meta-classifiers.



(a) One meta-class set (b) Two meta-class sets (c) Three meta-class sets

Figure 1: Simple illustration of decision boundaries (black solid lines) given by combinatorial classification with three seen classes, where three binary meta-classifiers are added one-by-one from (a) to (c). Unlike the standard classifier that creates decision boundaries for seen classes only, the combinatorial classification generates partitions using multiple coarse-grained classifiers, which are useful to discover novel classes.

- We introduce a contrastive learning approach to enhance the representations given by combinatorial embedding. We also present a transductive metric learning method based on pairwise pseudo-label estimation to learn the features of unlabeled data within the semantic structure identified by labeled ones.
- We demonstrate the outstanding performance of our model in the presence of novel classes through extensive evaluation on image retrieval and novel class discovery benchmarks.

2 Related work

This section first introduces recent approaches dealing with novel classes, which are roughly categorized into two groups: detection or rejection of unseen classes in test sets, and representation learning for classification and retrieval with unseen classes. Later in this section, we also discuss several related methods to the proposed algorithm in terms of methodology.

2.1 Learning in open-set setting

Departing from the closed world, a number of works recently consider the open-set setting, where novel classes appear during testing (Scheirer et al., 2012; Geng et al., 2020; Scheirer et al., 2014; Schölkopf et al., 2001; Bendale & Boult, 2016; Yoshihashi et al., 2019; Vareto et al., 2017; Júnior et al., 2017; Dan & Kevin, 2017). Early researches mainly focus on detecting out-of-distribution examples by learning binary classifiers (Scheirer et al., 2012; Schölkopf et al., 2001; Vareto et al., 2017; Dan & Kevin, 2017), or classifying the knowns while rejecting the unknowns (Scheirer et al., 2014; Bendale & Boult, 2016; Yoshihashi et al., 2019; Júnior et al., 2017). However, these approaches have significant challenges in distinguishing semantics between unseen classes; although some methods sidestep the issue by assigning rejected instances to new categories (Bendale & Boult, 2015; Shu et al., 2020; 2018), they require human intervention to annotate the rejected examples and consequently suffer from weak scalability.

To mitigate such limitations, transfer learning approaches have been proposed to model semantics between unseen classes. Using the representations learned from labeled data, the methods in this category perform clustering with unlabeled examples based on similarity prediction models (Hsu et al., 2018; 2019), ranking statistics (Han et al., 2020), and modified deep embedded clustering (Han et al., 2019) to capture their similarity and discrepancy. However, these approaches have two critical limitations. First, the problem setting is unrealistic because they assume that all unlabeled examples belong to unseen classes, and the number of novel classes is known in advance. Second, their main goal is to learn the representations of novel classes, which results in information loss about seen classes. On the contrary, we aim to discover novel categories without prior knowledge about unlabeled examples and learn the representations of all categories jointly, where unlabeled data may contain examples in both seen and unseen classes.

On the other hand, several hashing techniques (Zhang & Peng, 2017; Yan et al., 2017; Jin et al., 2020; Jang & Cho, 2020) learn approximate embeddings for image retrieval with both labeled and unlabeled data, which is generalizable to the examples in unseen classes. These methods focus on reducing quantization distortion in hash function by either entropy minimization (Zhang & Peng, 2017; Jang & Cho, 2020) or consistency regularization (Yan et al., 2017; Jin et al., 2020).

2.2 Combinatorial learning

Combinatorial learning framework reconstructs the solution space by the composition of the solutions from multiple heterogeneous tasks and there are several related approaches in this regard. Seo et al. (2018) formulates the image geolocalization problem as a classification task by combining multiple coarse-grained classifiers to reduce data deficiency and poor prediction granularity. A similar concept has been employed to learn noise-resistant classifiers (Seo et al., 2019) or recognize out-of-distribution examples (Vareto et al., 2017). Xuan et al. (2018) concatenate multiple representations learned on multiple class sets for metric learning.

Product quantization (Jegou et al., 2010; Ge et al., 2013), which is also related to combinatorial learning, constructs a large number of quantized regions given by a combination of subspace encodings to improve the performance of hash functions in an unsupervised manner. This approach is extended to learning quantization tables using image labels in (Jang & Cho, 2020; Klein & Wolf, 2019; Yu et al., 2018). However, they do not provide direct supervision for quantization and attempt to optimize the representation via the final classification loss, making the learned model suboptimal.

While all of these approaches are not studied in the presence of unlabeled examples during training except (Jang & Cho, 2020), the proposed algorithm leverages the composition of output representations for capturing the semantics of unlabeled data, which belong to either known or novel classes. Also, contrary to (Jang & Cho, 2020; Klein & Wolf, 2019; Yu et al., 2018), the proposed combinatorial embedding learns the representation with explicit supervision in the form of diverse meta-class labels and obtains a better embedding model for novel classes.

3 Proposed method

Suppose that we have a labeled dataset $\mathcal{D}_l = \{(x_i, y_i)\}_{i=1}^{N_l}$, where $x_i \in \mathbb{R}^d$ denotes an example and $y_i \in \mathcal{C}_l = \{c_1, \dots, c_K\}$ is its class label, and an unlabeled dataset $\mathcal{D}_u = \{(x_i)\}_{i=1}^{N_u}$ for training. Let $\mathcal{C}_l$ and $\mathcal{C}_u$ be the ground-truth class sets of the labeled and unlabeled data, respectively, where $\mathcal{C}_l \cap \mathcal{C}_u \neq \emptyset$ and $\mathcal{C}_l \neq \mathcal{C}_u$. We denote the novel class set by $\mathcal{C}_n = \mathcal{C}_u \backslash \mathcal{C}_l$. Our goal is to learn an unified model that is effective to represent novel classes as well as known ones by taking advantage of semantic relations across the two kinds of classes.

To this end, we propose a supervised combinatorial embedding approach followed by an unsupervised metric learning technique. For combinatorial embedding, we first construct multiple heterogeneous meta-class sets, each of which is obtained from different partitions of the base classes. We then obtain the combinatorial embedding vector of a base class by concatenating meta-class embeddings learned from the classifiers on the individual meta-class sets. Along with two supervised objectives on meta-class sets and base class set, we employ consistency regularization for robust combinatorial embeddings and transductive metric learning based on pairwise similarity for clustering the examples in both seen and unseen classes.

3.1 Preliminary: combinatorial classification

Before discussing the proposed framework, we first review combinatorial classification introduced in Seo et al. (2019). The main idea of the combinatorial classification is to construct a fine-grained classifier through a combination of multiple heterogeneous coarse-grained classifiers. Formally, given M coarse-grained classifiers $f^1, f^2, \dots, f^M$, defined over meta-class sets, we obtain a fine-grained combinatorial classifier $f \equiv f^1 \times f^2 \times \cdots \times f^M$, which is given by

$$\begin{array}{c} f^1 : x \in \mathbb{R}^d \rightarrow y \in \mathcal{C}^1 = \{c_1^1, \dots, c_{K_1}^1\} \\ \vdots \\ f^M : x \in \mathbb{R}^d \rightarrow y \in \mathcal{C}^M = \{c_1^M, \dots, c_{K_M}^M\} \end{array} \implies f : x \in \mathbb{R}^d \rightarrow y \in \mathcal{C}^1 \times \cdots \times \mathcal{C}^M. \tag{1}$$

The original paper (Seo et al., 2019) employs this idea to reduce label noise in datasets and learn noise-resistant classifiers. This paper adopts this concept to discover novel classes in datasets and learn their representations together with known ones.

3.2 Combinatorial embedding

Combinatorial embedding is a framework to learn a general representation, which embeds known and novel classes in a discriminative way using combinatorial classification.

We first construct M distinct partitions, denoted by $\mathcal{C}^m$ ($m = 1, \dots, M$). Each partition is referred to as a meta-class set, which has $K_m (\ll K)$ meta-classes, *i.e.*, $\mathcal{C}^m = \{c_1^m, \dots, c_{K_m}^m\}$, and each meta-class is typically constructed by a union of multiple base classes. Let an input image $x \in \mathbb{R}^d$ be mapped to a vector $z \in \mathbb{R}^{d_1}$ by a feature extractor $f_\theta(\cdot)$, *i.e.*, $z = f_\theta(x)$. The feature vector z is converted to M distinct vectors, $z^1, \dots, z^M$ ($z^m \in \mathbb{R}^{d_2}$), which are feature vectors for learning with meta-classifiers for individual meta-class sets, by applying M linear transform functions.

Unlike the combinatorial classification that estimates a class-conditional distribution, we estimate the embedding of each base class by the combination of multiple meta-class embeddings. Specifically, we construct M embedding heads with weight matrix $\Theta = \{\Theta^1, \cdots, \Theta^M\}$ ($\Theta^m \in \mathbb{R}^{d_2 \times K_m}$), and each head consists of meta-class representations for $\mathcal{C}^m$, denoted by $\Theta^m = \left[\theta_1^m, \cdots, \theta_{K_m}^m\right]$ ($\theta_k^m \in \mathbb{R}^{d_2}$).

The combinatorial embedding vector of a base class is obtained by a concatenation of the meta-class embeddings, which is formally given by $\pi = \left[\Phi(z^1, \Theta^1), \cdots, \Phi(z^M, \Theta^M)\right] \in \mathbb{R}^{d_2 M}$. Note that $\Phi(\cdot, \cdot)$ performs soft-assignments to individual meta-classes to enable backpropagation as

$$\Phi(z^m, \Theta^m) = \sum_{i=1}^{K_m} \frac{\exp\left(\lambda(z^m \cdot \theta_i^m)\right)}{\sum_{j=1}^{K_m} \exp\left(\lambda(z^m \cdot \theta_j^m)\right)} \theta_i^m, \tag{2}$$

where λ denotes a scaling factor. Feature vectors and embedding weights are ℓ_2-normalized before inner product to use cosine similarity as the distance metric. Note that the proposed embedding function enables us to analyze characteristics of unlabeled samples using their embedded vectors. For instance, examples in novel classes may share some meta-class labels with those of known classes while also having unique ones, which results in distinct feature vectors from the representations for the seen classes.

Using all the labeled examples, for which meta-class labels are also available, we propose the following two criteria to learn the discriminative representations.

Supervision on meta-class sets Our model learns the representations based on the meta-class labels using the normalized softmax loss (Zhai & Wu, 2019), which encourages a feature vector z^m to be close to the prototype of the ground-truth meta-class and located far away from the other meta-class prototypes. In detail, denoting by θ_+^m and θ_-^m the prototype of the ground-truth and non-ground-truth meta-class, respectively, the supervised objective on a meta-class set is defined as

$$\mathcal{L}_{\text{meta}} = -\sum_{m=1}^{M} \log \left(\frac{\exp\left(z^m \cdot \theta_+^m / \sigma\right)}{\sum_{\theta_-^i \in \Theta^m} \exp\left(z^i \cdot \theta_-^i / \sigma\right)} \right), \tag{3}$$

where each feature vector and prototype are ℓ_2-normalized before applying the softmax function to optimize with the cosine similarity, and σ denotes a temperature. Note that the meta-class embedding naturally introduces the inter-class relations into the model and leads better generalization for novel classes since the model learns the shared information from meta-class representations based on the examples in the multiple constituent base classes.

Supervision on base class set In addition to the meta-class level supervision, we also take advantage of the base class labels to preserve inter-class relations between known classes. To be specific, we train our model to the end that the combinatorial embedding vector π gets closer to the corresponding combinatorial ground-truth, $y^\pi = \left[\Theta^1[c^1], \dots, \Theta^M[c^M]\right]$, which is obtained by the concatenation of the ground-truth meta-class prototypes. As in the meta-class set case, let y_+^π be the prototype of the ground-truth class label, and y_-^π be the other prototypes. The supervised objective on the base class set is given by

$$\mathcal{L}_{\text{base}} = -\log \left(\frac{\exp\left(\pi \cdot y_+^\pi / \sigma\right)}{\sum_{y_-^\pi \in \mathcal{C}} \exp\left(\pi \cdot y_-^\pi / \sigma\right)} \right), \tag{4}$$

where π and $\{y_+^\pi, y_-^\pi\}$ are also ℓ_2-normalized and σ is a temperature parameter.

3.3 Combinatorial embedding with contrastive learning

To robustify the representations obtained by combinatorial embedding in the presence of novel classes, we perform the consistency regularization using both labeled and unlabeled examples. Among existing contrastive learning approaches (Chen et al., 2020; He et al., 2020; Chen & He, 2021), we adopt the SimSiam network (Chen & He, 2021) due to its simplicity and efficiency.

Let we get two combinatorial embeddings π_i and π'_i from the two randomly augmented views of an image x. A prediction MLP head (Grill et al., 2020), denoted as h, transforms the one to maximize the cosine similarity with respect to the other as

$$\mathcal{L}_{\text{cons}}\left(h(\pi), \pi'\right) = \frac{1}{n_l + n_u} \sum_{i}^{n_l+n_u} -\frac{h(\pi_i)}{\|h(\pi_i)\|_2} \cdot \frac{\pi'_i}{\|\pi'_i\|_2}, \tag{5}$$

where n_l and n_u are labeled and unlabeled examples in a batch, respectively, and $\|\cdot\|_2$ denotes ℓ_2-norm. Following Chen & He (2021), we do not backpropagate towards π'_i.

This loss function makes the examples in both seen and unseen classes transformation-invariant. Furthermore, it encourages the unseen class examples to be embedded on the proper locations in the joint embedding space, which improves the reliability of the pairwise pseudo-label estimation to be discussed in the next subsection.

3.4 Transductive metric learning via pairwise similarities

We identify the semantics of unlabeled data in the context of labeled ones using pairwise similarities given by metric learning with the estimated pseudo-labels. Since the class labels in $\mathcal{D}_u$ are unknown, we provide the relational supervision for the input feature vector pairs, (π_i, π_j), to learn the joint representations of labeled and unlabeled examples; the samples with similar features belong to the same class and have positive labels, *i.e.*, $\tilde{y}_{i,j} = 1$, while the dissimilar ones have negative ones, *i.e.*, $\tilde{y}_{i,j} = 0$. The question is how to estimate the pairwise pseudo-labels, $\tilde{y}_{i,j}$, for training.

To obtain the pairwise pseudo-ground-truths, $\tilde{y}_{i,j}$, we compute the similarities of combinatorial embedding vectors between each unlabeled example and the rest of images in a batch as follows:

$$\tilde{y}_{i,j} = \mathbb{I}\Big(\mathrm{s}\big(\pi_i, \pi_j\big) \geq \tau\Big), \tag{6}$$

where $\mathbb{I}(\cdot)$ denotes an indicator function and $\mathrm{s}(\cdot,\cdot)$ indicates an arbitrary similarity metric. Since unlabeled examples in the known classes are expected to be embedded properly and the novel classes are also embedded in the combinatorial feature space, the pseudo-label estimation based on equation 6 is sufficiently reliable in practice with a reasonable choice of the threshold, τ.

Once the pairwise labels are identified, we employ a metric learning strategy, which has been used for labeled data in (Jang & Cho, 2020), to learn the representations of novel class examples with respect to the knowns. The objective function is the following average cross-entropy loss:

$$\mathcal{L}_{\text{pair}} = \frac{1}{n_u} \sum_{i=1}^{n_u} \mathcal{L}_{\text{CE}}\left(s_i, \tilde{y}_{i,*}\right), \tag{7}$$

where $\boldsymbol{s}_i = [(z_i \cdot \pi_1), \ldots, (z_i \cdot \pi_{n_l+n_u})]$ and $\tilde{y}_{i,*} = \{0,1\}^{n_u+n_l}$ denote a cosine similarity vector and a pairwise pseudo-label vector between the i^{th} unlabeled feature vector and all combinatorial feature vectors in a batch. Since $\tilde{y}_{i,*}$ may not be a one-hot vector, we normalize the pseudo-ground-truth vector before applying the loss function. This loss function facilitates clustering novel class examples based on the similarities while maintaining the representations of the known classes given by equation 3 and equation 4.

3.5 Loss

The total loss is give by a weighted sum of four objective functions:

$$\mathcal{L} = \mathcal{L}_{\text{meta}} + \mathcal{L}_{\text{base}} + \alpha\mathcal{L}_{\text{cons}} + \beta\mathcal{L}_{\text{pair}}, \tag{8}$$

where α and β control the relative importance of the individual terms. The proposed framework performs a supervised classification, a self-supervised contrastive learning, and unsupervised metric learning jointly. The learned representations based on the proposed loss function should be effective for the examples in both known and novel classes.

Table 1: The mean Average Precision (mAP) for the different number of bits on CIFAR-10 and NUS-WIDE. The best mAP scores are in bold. GPQ with asterisk (*) is the result obtained by our reproduction from the TensorFlow implementation provided by the original authors.

Supervision	Method	CIFAR-10			NUS-WIDE		
		12 bits	24 bits	48 bits	12 bits	24 bits	48 bits
Unsupervised	OPQ (Ge et al., 2013)	0.107	0.119	0.138	0.341	0.358	0.373
	LOPQ (Kalantidis & Avrithis, 2014)	0.134	0.127	0.124	0.416	0.386	0.379
	ITQ (Gong et al., 2012)	0.157	0.165	0.201	0.488	0.493	0.503
Supervised	SDH (Shen et al., 2015)	0.185	0.193	0.213	0.471	0.490	0.507
	CNNH (Xia et al., 2014)	0.210	0.225	0.231	0.445	0.463	0.477
	NINH (Lai et al., 2015)	0.241	0.249	0.272	0.484	0.483	0.487
Supervised + Unlabeled data	SSDH (Zhang & Peng, 2017)	0.285	0.291	0.325	0.510	0.533	0.551
	SSGAH (Wang et al., 2018)	0.309	0.323	0.339	0.539	0.553	0.579
	GPQ* (Jang & Cho, 2020)	0.274	0.290	0.313	0.598	0.609	0.615
	SSAH (Jin et al., 2020)	0.338	0.370	0.379	0.569	0.571	0.596
	CombEmb (ours)	**0.670**	**0.731**	**0.750**	**0.704**	**0.724**	**0.727**

Table 2: mAP scores on CIFAR-100 and CUB-200 datasets with different number of bits.

Method	CIFAR-100			CUB-200		
	24 bits	48 bits	72 bits	24 bits	48 bits	72 bits
GPQ (Jang & Cho, 2020)	0.108	0.120	0.108	0.163	0.170	0.159
CombEmb (ours)	**0.144**	**0.186**	**0.209**	**0.263**	**0.296**	**0.286**

3.6 Image retrieval

The proposed approach based on combinatorial embedding (CombEmb) is evaluated in the image retrieval task, and we discuss asymmetric search algorithm for the image retrieval briefly.

Let z_q and z_b be the feature vectors of a query image x_q and a database image x_b, respectively. The proposed model, which is based on M partitions with K_m meta-classes per partition, requires $\sum_{m=1}^{M} \log_2(K_m)$ bits for storage to reconstruct the approximate representation of the database image x_b by $\bar{\pi}_b = \left[\Theta^1[c^1_{z^1_b}], \ldots, \Theta^M[c^M_{z^M_b}]\right]$, where $c^m_{z^m} \in \mathcal{C}^m$ denotes the matching meta-class of z^i_b. The distance between input query image and database image for asymmetric search is computed by the combination of the representations given by M partitions, which is given by

$$\sum_{m=1}^{M} \text{dist}(z_q^m, \bar{\pi}_b^m). \tag{9}$$

where dist$(\cdot, \cdot)$ denotes the cosine distance function and $\bar{\pi}_b^m$ means the matching meta class representation of the m^{th} partition.

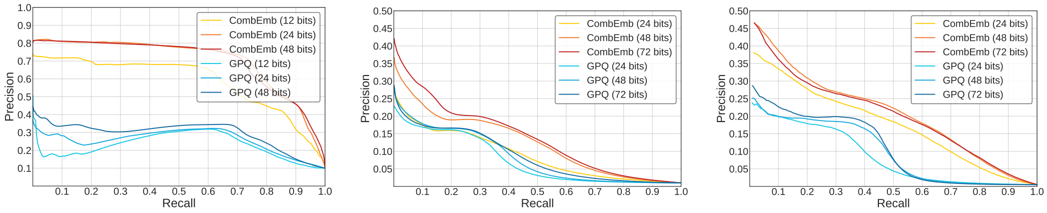
3.7 Discussion

The proposed algorithm provides a unique formulation for the representation learning of novel classes, which is given by the combination of meta-classifiers learned with the examples in known labels. The use of coarse-grained classifiers is helpful to capture common attributes across known and unknown classes, and the embeddings of the examples in novel classes are learned effectively by the combination of the multiple coarse-grained classifiers.

Our formulation is related to the concept of product quantization (Ge et al., 2013; Jegou et al., 2010) as discussed earlier. However, product quantization is originally proposed for unsupervised hashing, which simply maximizes the variance of data in subspace and enhances retrieval performance. Its extensions to supervised learning are limited to handling known classes only (Klein & Wolf, 2019) or fail to exploit the label information effectively for unseen class discovery (Jang & Cho, 2020).

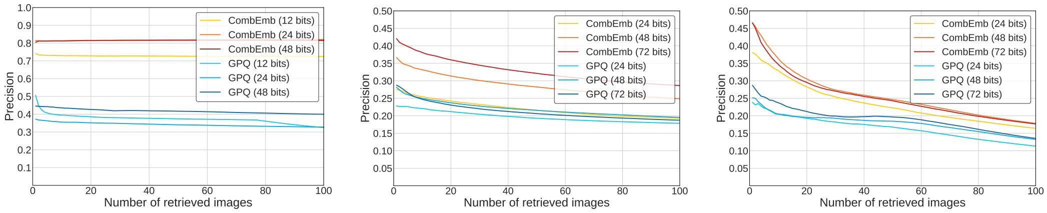
4 Experiments

This section presents the experimental results and the characteristics of the proposed approach in the application of image retrieval given a database containing both known and novel classes.



(a) CIFAR-10 (b) CIFAR-100 (c) CUB-200

Figure 2: Precision-Recall curves on the three tested datasets with various bit-lengths.



(a) CIFAR-10 (b) CIFAR-100 (c) CUB-200

Figure 3: Precision@100 curves on the three tested datasets with various bit-lengths.

4.1 Image retrieval of novel classes

Datasets and baselines We conduct experiments on four popular image retrieval benchmarks, CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009), NUS-WIDE (Chua et al., 2009), and CUB-200 (Welinder et al., 2010). For NUS-WIDE, we use the images associated with the 21 most frequent concepts, following (Liu et al., 2011). To simulate open-set environment in the datasets, we split the classes into two subsets—known (75%) and novel (25%) classes, which is identical to the protocol in (Sablayrolles et al., 2017). Specifically, 7, 15, 75, and 150 known classes are included in the labeled training datasets of CIFAR-10, NUS-WIDE, CIFAR-100, and CUB-200 respectively. Note that a training dataset contains unlabeled data, which may belong to either known or novel classes. For retrieval, the database contains the examples in both known and novel classes while the query set is composed of the ones in novel classes.

We compare our method, referred to as combinatorial embedding (CombEmb), with several image retrieval techniques based on hashing, which include OPQ (Ge et al., 2013), LOPQ (Kalantidis & Avrithis, 2014), ITQ (Gong et al., 2012). We also compare three supervised hashing methods: CNNH (Xia et al., 2014), NINH (Lai et al., 2015) and SDH (Shen et al., 2015), and four supervised hashing methods with additional unlabeled data: SSDH (Zhang & Peng, 2017), SSGAH (Wang et al., 2018), GPQ (Jang & Cho, 2020), SSAH (Jin et al., 2020).

Evaluation protocol Image retrieval performance is measured by the mean Average Precision (mAP). Since all compared methods are based on hashing, their capacities are expressed by bit-lengths; the capacity of CombEmb can be computed easily using the number of meta-classifiers and the number of meta-classes. We test four different bit-lengths (12, 24, 32, 48), and final results are given by the average of 4 different random splits.

Implementation details We develop our algorithm using Pytorch (Paszke et al., 2019). The backbone models are fine-tuned by AdamW (Loshchilov & Hutter, 2019) with weight decay factor 1×10^{-4}. For constituent classifiers, the number of meta-classes in each meta-class set (K_m) are fixed to 4, which are generated by the k-means clustering ($k = K_m$) of the seen class signatures, which are obtained by subspace projections of the fine-tuned weight vectors corresponding to the original classes in the classification layer. The number of meta-classifiers, M, is adjusted to match the bit-length of compared methods, and the dimensionality d_2 of z^m is set to 12.

Evaluation on benchmark datasets We first present the performance of the proposed approach, CombEmb, on CIFAR-10 and NUS-WIDE, in comparison to existing hashing-based methods. Ta-

Table 3: mAP scores on CIFAR-10 and CIFAR-100 datasets when we use the one half of classes as seen and the other half as novel. Our results on all datasets are averaged over 4 different class splits.

Method	CIFAR-10			CIFAR-100		
	12 bits	24 bits	48 bits	24 bits	48 bits	72 bits
GPQ (Jang & Cho, 2020)	0.245	0.245	0.261	0.100	0.095	0.099
CombEmb (ours)	**0.413**	**0.489**	**0.512**	**0.148**	**0.177**	**0.197**

Table 4: Performance of different pairwise pseudo-labeling methods on CIFAR-10 and NUS-WIDE.

Method	CIFAR-10			NUS-WIDE		
	12 bits	24 bits	48 bits	12 bits	24 bits	48 bits
k-means (MacQueen, 1967)	0.647	0.643	0.580	0.648	0.650	0.670
RankStats (Han et al., 2020)	0.592	0.485	0.488	0.675	0.651	0.684
CombEmb (ours)	**0.670**	**0.732**	**0.750**	**0.704**	**0.724**	**0.727**

ble 1 shows mAPs of all algorithms for four different bit-lengths for the representation of an instance, where the results of GPQ are from the reproduction on our data splits. CombEmb achieves the state-of-the-art performance in all cases on both datasets by significant margins. This is partly because, unlike previous hasing-based approaches which suffer from limited usage of unlabeled data other than quantization error reduction or consistency regularization, our models learn discriminative representations of unlabeled examples in novel classes by utilizing their inter-class relationships with labeled data through the combination of diverse meta-classifiers. In addition, the proposed metric learning technique further improves our embedding network via pseudo-labeling of pairwise similarities for unlabeled examples. The larger number of bits is effective for capturing the semantics in input images and achieving better performances in general. Figure 2 and 3 demonstrates more comprehensive results of CombEmb for image retrieval. Note that the area under the precision-recall curve corresponds to the mAP score.

We also apply our approach to more challenging datasets, CIFAR-100 and CUB-200, which contain fewer examples per class and potentially have troubles learning inter-class relations between seen and unseen classes. Table 2 presents that CombEmb outperforms GPQ consistently although the overall accuracies of both algorithms are lower than those on CIFAR-10 and NUS-WIDE. On the other hand, Table 3 shows that the proposed approach consistently outperforms GPQ with a fewer seen classes (50%) on CIFAR-10 and CIFAR-100. Note that the number of meta-classes in each meta-class set (K_m) for CIFAR-10 is set to 2 in this experiment.

Analysis on pairwise label estimation To understand the effectiveness of the pairwise labeling method proposed in equation 6, we compare it with the following two baselines: 1) using k-means clustering on the feature vectors to assign labels of unlabeled data points (k-means), and 2) adopting rank statistics (Han et al., 2020) between feature vectors before combinatorial embedding to estimate pairwise labels (RankStats). For the first baseline, we assume that the number of clusters is known and equal to the exact number of classes appearing in training. Table 4 implies that our pairwise label estimation strategy based on combinatorial embeddings outperforms other baselines.

Analysis of loss functions Table 5 demonstrates the contribution of individual loss terms in the experiment on CIFAR-10. Each of the four loss terms, especially the pairwise loss, turns out to be effective for improving accuracy consistently (except 12 bit-length case). Also, the consistency loss together with the pairwise loss is helpful to obtain the desirable tendency in accuracy with respect to bit-lengths. Overall, the consistency loss and the pairwise loss play complementary roles to achieve outstanding performance of our model.

Results with fewer labeled data Table 6 presents the performance on CIFAR-10 when 30% and 10% of the examples in the seen classes are labeled. Although the overall accuracy is degraded due to lack of supervision compared to the main experiment, the proposed algorithm outperforms GPQ by large margins regardless of bit-lengths.

4.2 Novel class discovery

To further validate the effectiveness of the proposed approach, we also compare with the state-of-the-art approaches for novel class discovery, including DTC (Han et al., 2019) and RankStats (Han

Table 5: Accuracy of the proposed approach with different combination of the loss terms.

$\mathcal{L}_{\text{meta}}$	$\mathcal{L}_{\text{base}}$	$\mathcal{L}_{\text{pair}}$	$\mathcal{L}_{\text{cons}}$	CIFAR-10		
				12 bits	24 bits	48 bits
✓				0.255	0.257	0.257
✓	✓			0.254	0.268	0.257
✓		✓		0.704	0.646	0.528
✓	✓	✓		**0.725**	0.670	0.532
✓	✓		✓	0.606	0.658	0.628
✓	✓	✓	✓	0.670	**0.731**	**0.750**

Table 6: mAP scores on CIFAR-10 with fewer labeled examples. We use 7 classes as seen classes.

Method	30% Labeled			10% Labeled		
	12 bits	24 bits	48 bits	12 bits	24 bits	48 bits
GPQ (Jang & Cho, 2020)	0.217	0.207	0.223	0.177	0.190	0.191
CombEmb (ours)	**0.463**	**0.607**	**0.713**	**0.227**	**0.254**	**0.378**

Table 7: Comparison with novel class discovery methods on CIFAR-10, CIFAR-100, and Tiny-ImageNet in terms of three evaluation metrics including ACC, NMI, and ARI.

Dataset	Method	ACC			NMI			ARI		
		Seen	Unseen	Total	Seen	Unseen	Total	Seen	Unseen	Total
CIFAR-10	DTC (Han et al., 2019)	51.5	57.5	42.3	48.9	24.7	44.2	34.5	24.2	27.9
	RankStats (Han et al., 2020)	72.8	83.6	66.1	73.6	49.6	68.6	62.4	33.2	51.7
	CombEmb (ours)	**90.0**	**89.0**	**88.2**	**78.8**	**64.7**	**78.3**	**80.5**	**76.8**	**77.3**
CIFAR-100	DTC (Han et al., 2019)	19.7	21.3	16.2	39.4	35.3	36.6	5.6	7.7	4.0
	RankStats (Han et al., 2020)	49.0	27.8	40.4	62.9	47.1	57.2	32.3	17.5	23.5
	CombEmb (ours)	**71.9**	**42.9**	**61.4**	**73.7**	**61.5**	**68.3**	**56.4**	**35.6**	**44.8**
Tiny-ImageNet	DTC (Han et al., 2019)	16.7	14.5	14.7	34.4	32.7	32.2	7.2	7.7	5.7
	RankStats (Han et al., 2020)	35.9	21.4	30.9	57.7	49.7	54.0	19.9	12.0	14.9
	CombEmb (ours)	**49.1**	**23.2**	**38.7**	**60.5**	**50.5**	**55.8**	**28.7**	**13.8**	**20.8**

et al., 2020), on CIFAR-10, CIFAR-100, and Tiny-ImageNet. This experiment adopts the same class split ratios of seen and unseen classes with image retrieval; the first 7, 75, and 150 classes in the three datasets are selected as seen classes and the half of their examples are labeled. For DTC and RankStat, we assume that the number of unseen classes is known and fine-tune the pretrained backbone model using the images in the seen classes before training with unlabeled data. We evaluate the performance using clustering accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI) using the predicted cluster indices. To report ACC, we solve the optimal assignment problem using Hungarian algorithm (Kuhn, 1955).

Table 7 presents the clustering performance of the learned representations by all the compared methods on CIFAR-10, CIFAR-100, and Tiny-ImageNet. CombEmb outperforms DTC and RankStat by significantly large margins for both seen classes and novel classes on all the datasets. Note that, unlike DTC and RankStats, we do not use any prior knowledge about the number of unseen classes. The results show that CombEmb learns effective representations for clustering in the presence of unseen classes in training datasets, which leads to the state-of-the-art performance in the novel class discovery task. We also provide t-SNE visualization of representations learned by CombEmb and discussion in comparison to DTC and RankStats in Appendix C.

5 Conclusion

This paper presents a novel representation learning approach, where only a subset of training examples are labeled while unlabeled examples may contain both known and novel classes. To address this problem, we proposed a combinatorial learning framework, which identifies and localizes the examples in unseen classes using the composition of the outputs from multiple coarse-grained classifiers on heterogeneous meta-class spaces. Our approach further strengthens the robustness of the representations through consistency regularization. We also utilized inter-class relations between seen classes for modeling the semantic structure with unseen classes and introduced a transductive metric learning strategy to estimate pairwise pseudo-labels for embedding unlabeled examples more effectively. The extensive experiments on the standard benchmarks for image retrieval and novel class discovery demonstrate the effectiveness of the proposed algorithm, and the various ablative studies show the robustness of our approach.

Reproducibility statement We provided implementation and evaluation details in Section 4.1 for image retrieval and Section 4.2 for novel class discovery. We submitted and will release the source code to facilitate the reproduction of our results.

REFERENCES

Abhijit Bendale and Terrance Boult. Towards open world recognition. In *CVPR*, 2015.

Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *CVPR*, 2016.

Maria Brbić, Marinka Zitnik, Sheng Wang, Angela O Pisco, Russ B Altman, Spyros Darmanis, and Jure Leskovec. MARS: discovering novel cell types across heterogeneous single-cell experiments. *Nature Methods*, 17(12):1200–1206, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021.

Haoang Chi, Feng Liu, Wenjing Yang, Long Lan, Tongliang Liu, Gang Niu, and Bo Han. Meta discovery: Learning to discover novel classes given very limited data. *arXiv preprint arXiv:2102.04002*, 2021.

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. NUS-WIDE: A real-world web image database from National University of Singapore. In *ACM-CIVR*, 2009.

Hendrycks Dan and Gimpel Kevin. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Tiezheng Ge, Kaiming He, Qifa Ke, and Jian Sun. Optimized product quantization for approximate nearest neighbor search. In *CVPR*, 2013.

Chuanxing Geng, Sheng-Jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020. doi: 10.1109/TPAMI.2020.2981604.

Yunchao Gong, Svetlana Lazebnik, Albert Gordo, and Florent Perronnin. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2916–2929, 2012.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020.

Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *ICCV*, 2019.

Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *ICLR*, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. In *ICLR*, 2018.

Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. Multi-class classification without multi-class labels. In *ICLR*, 2019.

Young Kyun Jang and Nam Ik Cho. Generalized product quantization network for semi-supervised image retrieval. In *CVPR*, 2020.

Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence*, 33(1):117–128, 2010.

Sheng Jin, Shangchen Zhou, Yao Liu, Chao Chen, Xiaoshuai Sun, Hongxun Yao, and Xian-Sheng Hua. SSAH: Semi-supervised adversarial deep hashing with self-paced hard sample generation. In *AAAI*, 2020.

Pedro R Mendes Júnior, Roberto M De Souza, Rafael de O Werneck, Bernardo V Stein, Daniel V Pazinato, Waldir R de Almeida, Otávio AB Penatti, Ricardo da S Torres, and Anderson Rocha. Nearest neighbors distance ratio open-set classifier. *Machine Learning*, 106(3):359–386, 2017.

Yannis Kalantidis and Yannis Avrithis. Locally optimized product quantization for approximate nearest neighbor search. In *CVPR*, 2014.

Benjamin Klein and Lior Wolf. End-to-end supervised product quantization for image search and retrieval. In *CVPR*, 2019.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Hanjiang Lai, Yan Pan, Ye Liu, and Shuicheng Yan. Simultaneous feature learning and hash coding with deep neural networks. In *CVPR*, 2015.

Wei Liu, Jun Wang, Sanjiv Kumar, and Shih-Fu Chang. Hashing with graphs. In *ICML*, 2011.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pp. 281–297. Oakland, CA, USA, 1967.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

Alexandre Sablayrolles, Matthijs Douze, Nicolas Usunier, and Hervé Jégou. How should we evaluate supervised hashing? In *ICASSP*, 2017.

Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7): 1757–1772, 2012.

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

Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

Paul Hongsuck Seo, Tobias Weyand, Jack Sim, and Bohyung Han. Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps. In *ECCV*, 2018.

Paul Hongsuck Seo, Geeho Kim, and Bohyung Han. Combinatorial inference against label noise. In *NeurIPS*, 2019.

Fumin Shen, Chunhua Shen, Wei Liu, and Heng Tao Shen. Supervised discrete hashing. In *CVPR*, 2015.

Yu Shu, Yemin Shi, Yaowei Wang, Yixiong Zou, Qingsheng Yuan, and Yonghong Tian. ODN: Opening the deep network for open-set action recognition. In *ICME*, 2018.

Yu Shu, Yemin Shi, Yaowei Wang, Tiejun Huang, and Yonghong Tian. P-ODN: Prototype-based open deep network for open set recognition. *Scientific reports*, 10(1):1–13, 2020.

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

Rafael Vareto, Samira Silva, Filipe Costa, and William Robson Schwartz. Towards open-set face recognition using hashing functions. In *IJCB*, 2017.

Guan'an Wang, Qinghao Hu, Jian Cheng, and Zengguang Hou. Semi-supervised generative adversarial hashing for image retrieval. In *ECCV*, 2018.

P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

Rongkai Xia, Yan Pan, Hanjiang Lai, Cong Liu, and Shuicheng Yan. Supervised hashing for image retrieval via image representation learning. In *AAAI*, 2014.

Hong Xuan, Richard Souvenir, and Robert Pless. Deep randomized ensembles for metric learning. In *ECCV*, 2018.

Xinyu Yan, Lijun Zhang, and Wu-Jun Li. Semi-supervised deep hashing with a bipartite graph. In *IJCAI*, 2017.

Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Naemura. Classification-reconstruction learning for open-set recognition. In *CVPR*, 2019.

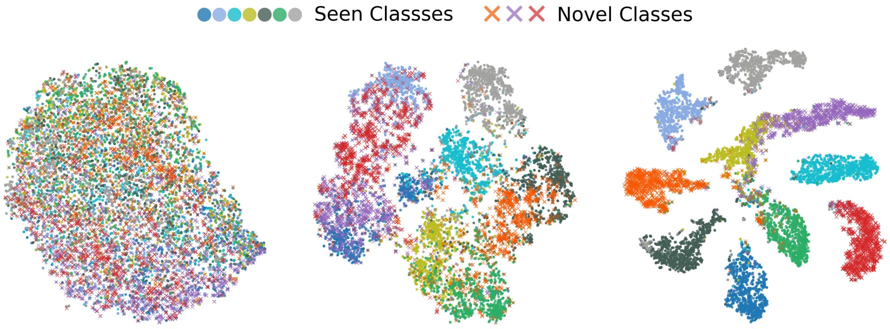
Tan Yu, Junsong Yuan, Chen Fang, and Hailin Jin. Product quantization network for fast image retrieval. In *ECCV*, 2018.

Andrew Zhai and Hao-Yu Wu. Classification is a strong baseline for deep metric learning. In *BMVC*, 2019.

Jian Zhang and Yuxin Peng. SSDH: Semi-supervised deep hashing for large scale image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(1):212–225, 2017.

A Evolution of the representation during training

Figure 4 illustrates the evolution of the learned representations of unlabeled data on CIFAR-10 using t-SNE (Van der Maaten & Hinton, 2008). Although the examples in different classes are mixed in the feature space at the beginning, they become more distinct from each other as training progresses, showing that our model effectively discovers novel categories without labels and learns their meaningful embeddings.

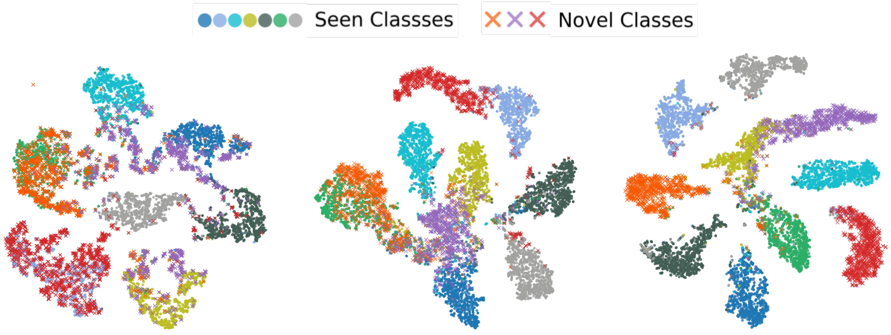


(a) Initial (b) Epoch-10 (c) Epoch-100

Figure 4: t-SNE visualization of the evolution of unlabeled data representations when we train the proposed model on CIFAR-10 with 7 seen classes and 3 novel classes.

B Visualization of embedding for image retrieval

Figure 5 visualizes the embeddings learned by GPQ (Jang & Cho, 2020), k-means, and the proposed method on CIFAR-10. According to the figure, GPQ fails to learn the distinct representation of novel classes; it tends to align the embedding of novel classes to that of the closest seen classes. The pseudo-labeling method given by k-means clustering with the oracle k also has troubles in learning the discriminative representation of the examples. To the contrary, we can observe that the proposed approach learns the representations, effectively discriminating both known and novel classes, through supervised combinatorial classification followed by unsupervised metric learning.

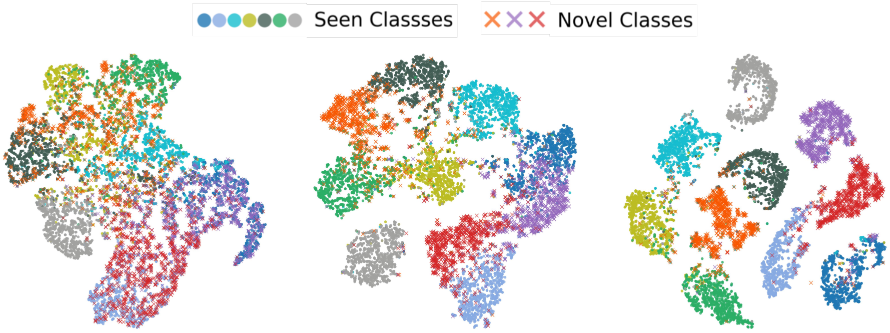


(a) GPQ (b) k-means (c) CombEmb (ours)

Figure 5: t-SNE visualization of CIFAR-10 using VGG, learned by GPQ, k-means and the proposed method. Visualization is based on 7 seen classes and 3 novel classes on CIFAR-10. Colors represent their ground-truth labels. Note that the proposed method embeds known and novel classes in a more discriminative way than other baselines. t-SNE hyperparameters are consistent in all three visualizations.

C VISUALIZATION OF EMBEDDING FOR NOVEL CLASS DISCOVERY

Figure 6 visualizes the embeddings learned by DTC Han et al. (2019), RankStats Han et al. (2020), and the proposed method on CIFAR-10. According to the figure, both DTC and RankStats have troubles in learning the discriminative representation between examples in seen classes and those in novel classes. In contrast, the proposed method embeds known and novel classes in a more discriminative way through supervised combinatorial classification followed by unsupervised metric learning.



(a) DTC (b) RankStats (c) CombEmb (ours)

Figure 6: t-SNE visualization of CIFAR-10 using VGG, learned by DTC, Rankstats and the proposed method. Visualization is based on 7 seen classes and 3 novel classes on CIFAR-10. Colors represent their ground-truth labels. t-SNE hyperparameters are consistent in all three visualizations.