Open-Set Representation Learning through Combinatorial Embedding

Geeho Kim[1] Junoh Kang[1] Bohyung Han[1,2]
Computer Vision Laboratory, ECE[1] & IPAI[2], Seoul National University
{snow1234, junoh.kang, bhhan}@snu.ac.kr

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 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 unsupervised pairwise relation learning. The proposed algorithm discovers novel concepts via a joint optimization for 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 on public datasets for image retrieval and image categorization with novel class discovery.

1. Introduction

Despite the remarkable success of machine learning fueled by deep neural networks, 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 [4, 5, 15–18], they often assume that all the unlabeled examples belong to unseen classes and/or the number of novel classes is known in advance, which makes their problem settings unrealistic.

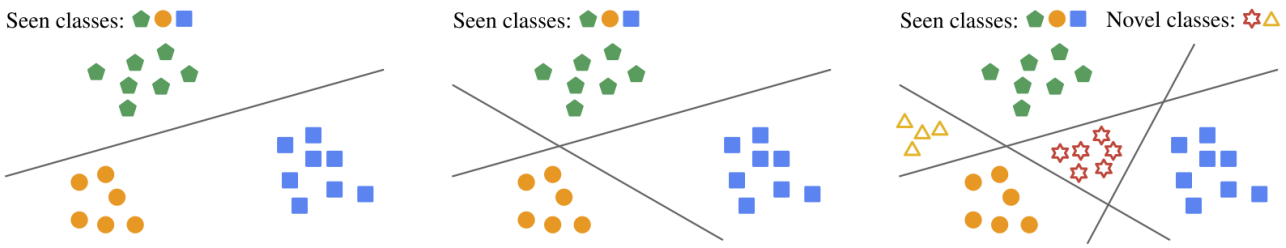
To address the limitations, this paper introduces an algorithm applicable to a more realistic setting. We aim to discover and learn the representations of unseen categories without any prior information or supervision about novel classes, where unlabeled data may contain examples in both seen and unseen classes. This task requires the model to be able to effectively identify unseen classes while preserving the information of previously seen classes. Our problem setting is more challenging than the case where the unlabeled data only consist of unseen classes because we have to solve an additional problem, predicting the membership of unlabeled examples between seen and unseen classes.

We propose a representation learning approach based on the concept of combinatorial classification [36], where the examples in unseen categories are identified by the composition of multiple meta-classifiers. Figure 1 illustrates the main idea of our *combinatorial embedding* framework, which forms partitions for novel classes via a combination of multiple classifiers for the meta-classes involving several constituent base classes. Images in the same meta-class potentially have common attributes that are helpful for knowledge transfer to novel classes, and we learn the representations of the images by the proposed combinatorial embedding. The learned representations via the combinatorial embedding become even stronger by unsupervised pairwise relation learning, which is effective to identify novel classes.

Our main contributions 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-class classifiers.
- We introduce an unsupervised learning approach to define pairwise relations, especially semantic structure between labeled and unlabeled examples, which further improves the quality of the representations given by combinatorial embedding.
- We demonstrate the outstanding performance of our model in the presence of novel classes through extensive evaluations on image retrieval and image categorization with novel class discovery benchmarks.

In the rest of this paper, we first review related works in Section 2 and discusses our main algorithm in Section 3. Section 4 presents our experimental results and Section 5 concludes this paper.



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

Figure 1. Conceptual 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 based on multiple coarse-grained classifiers creates and reserves partitions, which are distinct from those of seen classes, potentially corresponding to novel concepts.

2. Related work

This section first introduces recent approaches for open-set recognition, and then discusses several related methods to combinatorial learning.

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 [3, 9, 12, 22, 33–35, 41, 48]. Early researches mainly focus on detecting out-of-distribution examples by learning binary classifiers [9, 33, 35, 41], or classifying the knowns while rejecting the unknowns [3, 22, 34, 48]. 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 [2, 39, 40], 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 [17, 18], ranking statistics [15], and modified deep embedded clustering [16] to capture their similarity and discrepancy. However, these approaches have two critical limitations. First, the problem settings are 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. Recent works [5, 42] generalize the problem setting, where the unlabeled instances may come from both seen and novel classes. Cao *et al*. [5] revise the standard cross-entropy loss with an adaptive margin to prevent the model from being biased towards the seen classes while Vaze *et al*. [42] employ two contrastive losses to pretrain the representations and adopt k-means++ clustering [1] for evaluation. However, these approaches still require prior information about the number of novel classes or computationally expensive modules to estimate the number of novel classes. On the contrary, we do not use any information about the number of novel classes for training because we discover novel categories based on the outputs from the meta-classifiers.

On the other hand, several hashing techniques [19, 21, 47, 51] learn approximated embeddings for image retrieval with both labeled and unlabeled data, which is generalizable to the examples in unseen classes. They focus on reducing quantization distortion in hash function by either entropy minimization [19, 51] or consistency regularization [21, 47].

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*. [37] formulate 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 [36] or recognize out-of-distribution examples [41]. Xuan *et al*. [46] concatenate multiple representations learned on multiple class sets for metric learning.

Product quantization [11, 20], 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 [19, 25, 49]. However, they do not provide direct supervision for quantization but 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 GPQ [19], the proposed algorithm leverages the composi-

tion of output representations for capturing the semantics of unlabeled data, which belong to either known or novel classes. Also, contrary to [19, 25, 49], 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 Approach

Suppose that we are given a labeled dataset, $\mathcal{D}_l = \{(x_i, y_i)\}_{i=1}^{N_l}$, where $x_i \in \mathbb{R}^d$ denotes an input example and $y_i \in \mathcal{C}_l = \{c_1, \dots, c_K\}$ is its class label, as well as 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 and two unsupervised pairwise learning techniques. For combinatorial embedding, we first construct multiple heterogeneous meta-class sets, each of which is obtained from a unique partition of the base classes. We then obtain the combinatorial embedding vector of a base class by concatenating meta-class embeddings learned from the classifiers over the individual meta-class sets. Along with the supervised learning objective, we also perform unsupervised learning based on contrastive loss and consistency regularization for understanding pairwise relations of both seen and unseen classes.

3.1. Supervised Combinatorial Embedding

The main idea of the supervised combinatorial embedding is to learn the general representations, which embed known and novel classes in a discriminative way, using a composition of multiple heterogeneous coarse-grained classifiers corresponding to meta-class sets. Formally, if we are given M coarse-grained classifiers $f^1, f^2, \dots, f^M$, defined over meta-class sets as

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

we obtain a fine-grained combinatorial classifier $f \equiv f^1 \times f^2 \times \cdots \times f^M$, which is given by

$$f : x \in \mathbb{R}^d \rightarrow y \in \mathcal{C}^1 \times \cdots \times \mathcal{C}^M. \tag{2}$$

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 partitioned to M distinct subvectors, $z^1, \dots, z^M$ ($z^m \in \mathbb{R}^{d_2}, d_1 = Md_2$), which are feature vectors for learning the meta-classifiers for the corresponding meta-class sets. We estimate the embedding of each base class based on the meta-class embeddings in the meta-classifiers. 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 corresponds to a classifier for a meta-class set $\mathcal{C}^m$ whose parameters consist of the prototypes of the meta-classes, 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 of a base class is given by a concatenation of the meta-class embeddings, which is formally given by $\pi(z; \Theta) = \left[\Phi(z^1, \Theta^1), \cdots, \Phi(z^M, \Theta^M)\right] \in \mathbb{R}^{d_2 M}$. Note that $\Phi(\cdot, \cdot)$ performs the soft assignment [49] of z^m to individual meta-classes to enable backpropagation as

$$\Phi(z^m, \Theta^m) = \sum_{i=1}^{K_m} \frac{\exp\left(\lambda(z^m \cdot \theta_i^m)\right)}{\sum_{j=1}^{K_m} \exp\left(\lambda(z^m \cdot \theta_j^m)\right)} \theta_i^m, \tag{3}$$

where λ is a sufficiently large scaling factor to approximate the function to a discrete argmax function. Feature vectors and embedding weights are ℓ_2-normalized before the inner product to use cosine similarity as the distance metric. Note that the proposed embedding function enables us to characterize the semantics of unlabeled samples using their embeddings. For instance, supposing that an example in a novel class has the same meta-class label as those in some known classes with a meta-class set while having a different one in another meta-class set, we can compute the unique embedding of the novel class with respect to those of the seen classes.

Using all the labeled examples, for which meta-class labels are also available, our model learns the representations based on the meta-class labels using the normalized softmax loss [50], which encourages a feature vector z^m to be close to the prototype of the ground-truth meta-class and far away from the other meta-class prototypes. Formally, denoting by θ_+^m the prototype of the ground-truth, the supervised objective on a meta-class set is defined as

$$\mathcal{L}_{\text{meta}} = -\sum_{m=1}^{M} \log\left(\frac{\exp\left(z^m \cdot \theta_+^m / \tau\right)}{\sum_{\theta^i \in \Theta^m} \exp\left(z^m \cdot \theta^i / \tau\right)}\right), \tag{4}$$

where each feature vector and prototype are ℓ_2-normalized and τ represents a temperature of the softmax function. 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 using the examples in the multiple constituent base classes.

3.2. Unsupervised Learning of Pairwise Relations

The combinatorial embedding is obtained by a large number of partitions through the composition of many meta-classifiers while it tends to scatter unlabeled examples over the feature space. To learn the proper embeddings of unlabeled samples, especially the ones in novel classes, we consider two kinds of pairwise relations; one is the psuedo-label consistency between two instances and the other is the representation consistency between two augmented examples of an image. These two objectives are learned in an unsupervised way based on the combinatorial embeddings of images as follows.

Contrastive learning for pseudo-label consistency We capture the semantics of unlabeled data in the context of labeled ones and perform pairwise pseudo-label estimation based on similarities between two real examples. Since the class labels in $\mathcal{D}_u$ are unknown, we provide the relational supervision for each input feature vector pair, (z_i, z_j), to learn the representations properly for both labeled and unlabeled examples. To this end, the examples with similar features are assumed to belong to the same class and regarded as a positive pair via the following procedure.

We leverage the labeled dataset $\mathcal{D}_l$ to bootstrap representations and use classification outputs from meta-classifiers Θ to infer relationships between examples. Specifically, the positive examples are selected based on the similarities of combinatorial embedding vectors between each unlabeled example and the rest of the images in a batch, which is given by

$$\mathcal{P}_z = \{\tilde{z} | \tilde{z} \in B_l \cup B_u, \pi(z;\Theta) \cdot \pi(\tilde{z};\Theta) \geq \gamma)\}, \quad (5)$$

where $\mathcal{B}_l$ and $\mathcal{B}_u$ are sets of feature vectors corresponding to labeled and unlabeled examples in the current mini-batch. Since unlabeled examples in known classes typically yield good representations thanks to labeled counterparts in the same class and the novel classes can be embedded properly in the combinatorial feature space, we expect the pseudo-label estimation by (5) is sufficiently reliable in practice under a reasonable choice of the threshold, γ.

Once the positive pairs are identified, we employ a contrastive loss [24] to enforce the similarity of the positive pairs as

$$\mathcal{L}_{\text{sim}}(z) = -\frac{1}{|\mathcal{P}_z|} \sum_{z_+ \in \mathcal{P}_z} \log \frac{\exp\left(z \cdot \pi(z_+;\Theta)\right)}{\sum_{\tilde{z} \in \mathcal{B}_l \cup \mathcal{B}_u} \exp\left(z \cdot \pi(\tilde{z};\Theta)\right)}, \quad (6)$$

where z and $\pi(\cdot;\Theta)$ are also ℓ_2-normalized. This loss term facilitates clustering novel class examples based on the cosine similarity while maintaining the representations of known class data given by (4). It also allows us to jointly learn the deep feature representations in both the original space and the combinatorial embedding space.

Consistency regularization of combinatorial embedding Besides the label consistency between two different examples, we perform the consistency regularization with both labeled and unlabeled data to robustify the representations obtained by combinatorial embedding in the presence of novel classes. Given two feature vectors z and z' for the two augmented views of an image $x \in \mathcal{D}_l \cup \mathcal{D}_u$, we minimize the negative cosine similarity between their combinatorial embeddings as

$$\mathcal{L}_{\text{cons}}(z, z') = -\frac{h(\pi(z;\Theta))}{\|h(\pi(z;\Theta))\|_2} \cdot \frac{\pi(z';\Theta)}{\|\pi(z';\Theta)\|_2}, \quad (7)$$

where $h(\cdot)$ denotes a prediction head [14], and $\|\cdot\|_2$ denotes ℓ_2-norm. Following [7], we do not backpropagate through $\pi(z';\Theta)$. This loss encourages the examples of both seen and unseen classes to be embedded in the proper locations within the common embedding space, which improves the reliability of the positive pair estimation.

3.3. Loss

The total loss is a weighted sum of the three objective functions as

$$\mathcal{L} = \mathcal{L}_{\text{meta}} + \alpha \mathcal{L}_{\text{sim}} + \beta \mathcal{L}_{\text{cons}}, \quad (8)$$

where α and β control the relative importance of the individual terms. The proposed framework jointly performs a supervised classification and two unsupervised pairwise relation learnings. The learned representations based on the proposed loss function should be effective for the examples in both known and novel classes.

3.4. 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.

Our formulation is related to the concept of product quantization (PQ) [11, 20] as discussed earlier. However, PQ is originally proposed for unsupervised hashing, which simply maximizes the variances of data in multiple subspaces and enhances retrieval performance in terms of accuracy and speed. Its extensions to supervised learning are limited to handling known classes only [25] or fail to exploit the label information effectively for learning the representations of unlabeled novel class examples [19].

Table 1. The mean Average Precision (mAP) for different bit-lengths on CIFAR-10 and NUS-WIDE. The best mAP scores are in bold. GPQ with an asterisk (*) presents the result from our reproduction with the 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 [11]	0.107	0.119	0.138	0.341	0.358	0.373
	LOPQ [23]	0.134	0.127	0.124	0.416	0.386	0.379
	ITQ [13]	0.157	0.165	0.201	0.488	0.493	0.503
Supervised	SDH [38]	0.185	0.193	0.213	0.471	0.490	0.507
	CNNH [45]	0.210	0.225	0.231	0.445	0.463	0.477
	NINH [29]	0.241	0.249	0.272	0.484	0.483	0.487
Supervised + Unlabeled data	SSDH [51]	0.285	0.291	0.325	0.510	0.533	0.551
	SSGAH [43]	0.309	0.323	0.339	0.539	0.553	0.579
	GPQ* [19]	0.274	0.290	0.313	0.598	0.609	0.615
	SSAH [21]	0.338	0.370	0.379	0.569	0.571	0.596
	CombEmb (ours)	**0.667**	**0.692**	**0.720**	**0.687**	**0.693**	**0.706**

4. Experiments

This section presents the experimental results and the characteristics of our method in the applications of image retrieval and novel class discovery given a database composed of both known and novel classes.

4.1. Image Retrieval with Novel Class Examples

Image retrieval is the task to identify images that belong to the same class given a query, where the database contains the examples in both known and novel classes. In our scenarios, query images are sampled from novel classes.

Image retrieval using combinatorial embedding We discuss an asymmetric search algorithm for image retrieval based on combinatorial embedding. 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 to store the approximate representation of the database image x_b, denoted by $\bar{z}_b = \left[\Theta^1[c^1_{z^1_b}], \ldots, \Theta^M[c^M_{z^M_b}]\right]$, where $c^m_{z^m_b} \in \mathcal{C}^m$ indicates the meta-class label of z^m_b. The distance between input query image and database image for asymmetric search is computed by the combination of the representations with M partitions, which is given by

$$\sum_{m=1}^{M} \text{dist}(z^m_q, \bar{z}^m_b). \tag{9}$$

where $\text{dist}(\cdot,\cdot)$ is the cosine distance function and $\bar{z}^m_b$ is the matching meta-class representation of the m^{th} partition.

Datasets We conduct experiments on four popular image retrieval benchmarks, CIFAR-10 [26], CIFAR-100 [26], NUS-WIDE [8], and CUB-200 [44]. For NUS-WIDE, we use the images associated with the 21 most frequent concepts, following [30]. To simulate an open-set environment in the datasets, we split the classes into two subsets, known (75%) and novel (25%) classes, and set the half of the examples in known classes as labeled, which is identical to the protocol in [32]. 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.

Baselines We compare the proposed approach, referred to as combinatorial embedding (CombEmb), with several image retrieval baselines based on hashing, which include OPQ [11], LOPQ [23], and ITQ [13]. We also compare three supervised hashing techniques including CNNH [45], NINH [29], and SDH [38], and four supervised hashing methods with additional unlabeled data such as SSDH [51], SSGAH [43], GPQ [19], and SSAH [21]. We extract feature descriptors from AlexNet [27] pretrained on ImageNet [10] for all the methods except GPQ [19], which adopts the modified VGG network for CIFAR-10/100 and AlexNet for NUS-WIDE as feature extractors.

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 three different bit-lengths $\{12, 24, 48\}$, and final results are given by the average of 4 different class splits.

Implementation details The backbone models and the embedding heads are fine-tuned by AdamW [31] with a weight decay factor of 1×10^{-4}. For meta-classifiers,

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

Dataset	Method	24 bits	48 bits	72 bits
CIFAR-100	GPQ	0.108	0.120	0.112
	CombEmb (ours)	**0.154**	**0.188**	**0.208**
CUB-200	GPQ	0.167	0.184	0.192
	CombEmb (ours)	**0.304**	**0.337**	**0.336**

Table 3. mAP scores on CIFAR-10 and CIFAR-100 with 50% of seen classes and another 50% of novel ones.

Dataset	Method	24 bits	48 bits
CIFAR-10	GPQ	0.231	0.245
	CombEmb (ours)	**0.448**	**0.491**
CIFAR-100	GPQ	0.104	0.117
	CombEmb (ours)	**0.165**	**0.179**

the number of meta-classes in each meta-class set (K_m) is fixed to 4 to simplify the experiment. 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. For meta-class set configuration, we generate M meta-class sets by iteratively performing k-means clustering ($k = K_m$) over class embeddings. We obtain the class embeddings from the classification weight vectors pretrained on labeled data. To ensure diverse meta-class sets, we randomly sample $Q(\ll d_1)$-dimensional subspaces of the class embeddings for each meta-class set generation. We list the implementation details of the proposed method and the compared algorithms in the supplementary document.

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. Tab. 1 shows mAPs of all algorithms for three different bit-lengths, where the results of GPQ are from the reproduction on our data splits. CombEmb achieves state-of-the-art performance in all cases on both datasets by significant margins. This is partly because, unlike previous hashing-based approaches that suffer from limited usage of unlabeled data other than quantization error reduction or consistency regularization, our model learns 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 unsupervised pairwise relation learning further improves our embedding network via enforcing similarities between unlabeled examples and their pseudo-positives. The larger number of bits is effective for capturing the semantics in input images and achieving better performances in general.

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

Dataset	Method	12 bits	24 bits	48 bits
CIFAR-10	k-means	0.529	0.593	0.510
	RankStats [15]	0.572	0.635	0.552
	CombEmb (ours)	**0.667**	**0.692**	**0.720**
NUS-WIDE	k-means	0.652	0.638	0.640
	RankStats [15]	0.641	0.649	0.656
	CombEmb (ours)	**0.687**	**0.693**	**0.706**

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

$\mathcal{L}_{\text{meta}}$	$\mathcal{L}_{\text{sim}}$	$\mathcal{L}_{\text{cons}}$	CIFAR-10		
			12 bits	24 bits	48 bits
✓			0.252	0.253	0.266
✓		✓	0.510	0.596	0.623
✓	✓		**0.687**	0.676	0.619
✓	✓	✓	0.667	**0.692**	**0.720**

We also apply CombEmb 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. Tab. 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, Tab. 3 shows that CombEmb consistently outperforms GPQ with a fewer seen classes (50%) on CIFAR-10 and CIFAR-100. In this experiment, K_m for CIFAR-10 is set to 2 since we have only 5 seen classes in CIFAR-10.

Analysis on pairwise label estimation To understand the effectiveness of the positive pair estimation proposed in (5), we compare the strategy with the following two baselines: 1) using k-means clustering on the feature vectors to assign labels of unlabeled data (k-means), and 2) adopting rank statistics [15] between feature vectors in the original space to estimate pairwise labels (RankStats). For the first baseline, we assume the ideal case in which the number of clusters is known and equal to the exact number of classes appearing in training. Tab. 4 implies that our label estimation strategy based on combinatorial embeddings outperforms other baselines.

Analysis of loss functions Tab. 5 demonstrates the contribution of individual loss terms on CIFAR-10. Each of the three loss terms, especially the similarity loss ($\mathcal{L}_{\text{sim}}$) and consistency loss ($\mathcal{L}_{\text{cons}}$), turn out to be effective for improving accuracy consistently. Also, the similarity loss together with the consistency loss is helpful to obtain the desirable tendency in accuracy with respect to bit-lengths. Note that

Table 6. Sensitivity of CombEmb in mAP to the number of meta-classes K_m, where the bit-length is controlled by adjusting the number of meta-classifiers M given K_m.

K_m	CIFAR-100			CUB200		
	24 bits	48 bits	72 bits	24 bits	48 bits	72 bits
2	0.153	0.175	0.179	0.296	0.330	0.334
4	0.154	0.188	0.208	0.304	0.337	0.336
8	0.141	0.196	0.223	0.291	0.337	0.338

Table 7. mAP scores on CIFAR-10 with fewer labeled examples of seen classes, where 7 classes are set as seen classes.

Ratio	Method	12 bits	24 bits	48 bits
30% labeled	GPQ	0.217	0.207	0.223
	CombEmb (ours)	**0.354**	**0.572**	**0.697**
10% labeled	GPQ	0.177	0.190	0.191
	CombEmb (ours)	**0.184**	**0.264**	**0.498**

the proposed combinatorial embedding learns basic representations suitable for both seen and unseen examples and the unsupervised pairwise relation learning improves performance dramatically on top of them.

Analysis on the number of meta-classes Tab. 6 presents the performance of CombEmb by varying the number of the meta-classes K_m while the bit-lengths are controlled as the same values by adjusting M, the number of meta-classifiers. We observe that larger bit-lengths are consistently helpful for improving the accuracy of CombEmb while K_m and M alone have limited impacts on performance.

Results with fewer labeled data We perform experiments when 30% and 10% of the examples in the seen classes are labeled and present the results in Tab. 7. These settings are more realistic and challenging than the environment of our main experiments. Although the overall accuracy is degraded compared to the main results due to the lack of supervision, the proposed algorithm outperforms GPQ by large margins regardless of bit-lengths. Note that when we use large bit-lengths (48 bits), the performance gap becomes more significant than the experiments in Tab. 1, achieving $3.1\times$ and $2.6\times$ accuracy gains with 30% and 10% of labeled seen-class examples, respectively.

4.2. Categorization with Novel Class Discovery

We evaluate the performance of CombEmb on image categorization with novel class discovery. The goal of this task is to cluster unlabeled examples that belong to either seen or novel classes into a predefined number of groups based on their semantic relations. This task is more natural and challenging than the standard novel class discovery that only considers unseen classes in unlabeled data.

Datasets We evaluate the proposed approaches on three standard datasets including CIFAR-10, CIFAR-100, and Tiny-ImageNet. Similar to the experiments for image retrieval, we split the classes into 75% seen and 25% novel classes: the first 7, 75, and 150 classes in CIFAR-10, CIFAR-100, and Tiny-ImageNet are respectively selected as seen classes. Following [5, 42], we assume that the half of examples in seen classes are labeled while setting the rest in seen classes and examples in novel classes as unlabeled. Note that the unlabeled data may belong to either known or novel classes.

Baselines We compare CombEmb with the state-of-the-art approaches in novel class discovery including DTC [16], RankStats [15], NCL [53], and DualRank [52]. We additionally consider two more methods in a similar setting: ORCA [5] and GCD [42]. Since the classification heads for unlabeled data in DTC, RankStats, NCL, and DualRank cannot handle seen classes, we increase the dimensionality of the classifiers to the total number of classes in datasets. This extension requires estimating the number of novel classes [16,42] unless it is given in advance. For all datasets, we use ResNet-18 as a backbone and pretrain all the compared methods with SimCLR [6] while DTC, RankStats, and NCL additionally fine-tune their models with the labeled data. We describe the implementation details of all algorithms in the supplementary document. For evaluation, we first identify the cluster membership of each example in the test set via k-means clustering, and then compute clustering accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI) using the cluster indices. Note that, to report ACC, we solve the optimal assignment problem using the Hungarian algorithm [28].

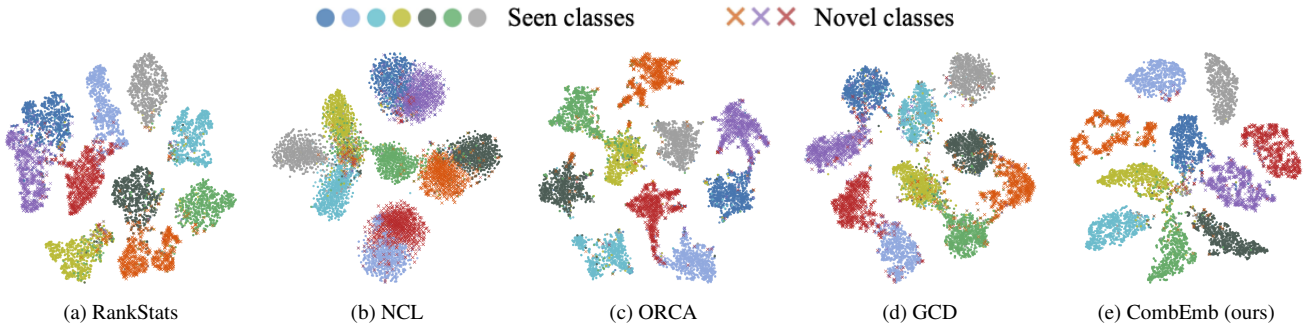
Results Tab. 8 presents the clustering performance of the learned representations by all the compared methods on all the three datasets. CombEmb outperforms the baselines for both seen classes and novel classes in most cases. The results show that CombEmb learns effective representations for clustering in the presence of unseen classes in training datasets, which leads to state-of-the-art performance. Figure 2 visualizes the embeddings learned by RankStats, NCL, ORCA, GCD, and CombEmb on CIFAR-10. Our method embeds known and novel classes in a more discriminative way through supervised combinatorial classification followed by unsupervised learning, while other methods suffer from learning the discriminative representations, especially between the novel classes and their closest seen classes.

5. Conclusion

This paper presents a novel representation learning approach, where only a subset of training examples are la-

Table 8. Comparison with novel class discovery methods on CIFAR-10, CIFAR-100, and Tiny-ImageNet in terms of ACC, NMI, and ARI. Dagger (†) denotes that the dimensionality of the classifier in the original method is extended to the total number of classes in the dataset.

Dataset	Method	ACC			NMI			ARI		
		Seen	Unseen	Total	Seen	Unseen	Total	Seen	Unseen	Total
CIFAR-10	DTC† [16]	77.44	72.00	75.81	62.91	65.82	63.25	61.98	65.33	56.89
	RankStats† [15]	82.22	88.23	84.01	72.03	75.17	73.65	66.68	79.54	67.08
	NCL† [53]	78.47	85.23	80.50	69.09	74.97	66.18	68.06	78.35	56.56
	DualRank† [52]	83.89	86.91	84.79	72.97	74.37	74.98	67.61	78.63	68.02
	ORCA [5]	87.70	92.83	89.24	74.34	**81.81**	78.21	76.46	**88.13**	78.12
	GCD [42]	88.82	88.23	88.63	76.28	75.53	77.99	77.52	81.19	76.75
	CombEmb (ours)	**89.02**	**92.96**	**89.98**	**77.97**	80.43	**79.83**	**80.26**	85.63	**79.19**
CIFAR-100	DTC† [16]	42.67	27.44	38.86	54.71	44.77	50.04	20.27	23.84	15.24
	RankStats† [15]	47.33	34.79	42.49	62.80	50.56	58.29	30.93	21.05	22.63
	NCL† [53]	53.13	35.80	48.80	63.23	54.24	58.56	38.93	27.68	30.06
	DualRank† [52]	46.08	36.47	42.54	61.59	52.20	57.57	29.83	24.09	23.00
	ORCA [5]	64.85	44.83	56.77	66.18	58.89	62.64	46.22	38.68	38.06
	GCD [42]	66.04	38.53	55.59	69.61	57.67	64.56	49.99	30.38	38.85
	CombEmb (ours)	**69.19**	**51.41**	**62.11**	**70.77**	**63.81**	**67.71**	**52.22**	**43.37**	**43.37**
Tiny-ImageNet	DTC† [16]	16.76	14.47	16.19	34.41	32.70	32.19	7.23	7.72	5.72
	RankStats† [15]	31.49	19.75	26.87	54.76	47.52	51.14	16.16	9.86	11.87
	NCL† [53]	35.27	18.90	31.18	55.08	47.16	51.10	17.89	10.01	12.85
	DualRank† [52]	29.76	18.93	26.70	53.86	48.03	50.53	14.88	9.69	11.08
	ORCA [5]	47.46	22.55	38.58	60.26	49.08	55.23	27.62	13.37	19.72
	GCD [42]	45.77	20.59	38.03	61.58	49.74	56.54	27.21	11.45	19.32
	CombEmb (ours)	**53.76**	**27.41**	**43.73**	**63.91**	**54.07**	**59.01**	**33.43**	**17.70**	**23.40**



(a) RankStats (b) NCL (c) ORCA (d) GCD (e) CombEmb (ours)

Figure 2. t-SNE visualization for the data embedding of CIFAR-10, learned by Rankstats, NCL, ORCA, GCD, and CombEmb. The visualization is based on 7 seen classes and 3 novel classes. Colors represent their ground-truth labels.

beled 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 improves the semantic structures and the robustness of the representations via unsupervised relation learning. The extensive experiments on the standard benchmarks for image retrieval and image categorization with novel class discovery demonstrate the effectiveness of the proposed algorithm, and the various ablative studies show the robustness of our approach.

Acknowledgments This work was partly supported by Samsung Advanced Institute of Technology (SAIT) and the NRF Korea grant [No. 2022R1A2C3012210, Knowledge Composition via Task-Distributed Federated Learning; No.2022R1A5A708390811, Trustworthy Artificial Intelligence]. It is also supported in part by the IITP grants [2021-0-02068; 2021-0-01343] funded by the Korea government (MSIT).

References

[1] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *ACM-SIAM symposium on Discrete algorithms*, 2007. 2

[2] Abhijit Bendale and Terrance Boult. Towards open world recognition. In *CVPR*, 2015. 2

[3] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *CVPR*, 2016. 2

[4] 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. 1

[5] Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. In *ICLR*, 2022. 1, 2, 7, 8

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 7

[7] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021. 4

[8] 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. 5

[9] Hendrycks Dan and Gimpel Kevin. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017. 2

[10] 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. 5

[11] Tiezheng Ge, Kaiming He, Qifa Ke, and Jian Sun. Optimized product quantization for approximate nearest neighbor search. In *CVPR*, 2013. 2, 4, 5

[12] Chuanxing Geng, Sheng-Jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 2

[13] 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. 5

[14] 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. 4

[15] 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. 1, 2, 6, 7, 8

[16] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *ICCV*, 2019. 1, 2, 7, 8

[17] Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. In *ICLR*, 2018. 1, 2

[18] Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. Multi-class classification without multi-class labels. In *ICLR*, 2019. 1, 2

[19] Young Kyun Jang and Nam Ik Cho. Generalized product quantization network for semi-supervised image retrieval. In *CVPR*, 2020. 2, 3, 4, 5

[20] 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. 2, 4

[21] 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. 2, 5

[22] 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. 2

[23] Yannis Kalantidis and Yannis Avrithis. Locally optimized product quantization for approximate nearest neighbor search. In *CVPR*, 2014. 5

[24] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020. 4

[25] Benjamin Klein and Lior Wolf. End-to-end supervised product quantization for image search and retrieval. In *CVPR*, 2019. 2, 3, 4

[26] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Technical Report*, 2009. 5

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 5

[28] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 7

[29] Hanjiang Lai, Yan Pan, Ye Liu, and Shuicheng Yan. Simultaneous feature learning and hash coding with deep neural networks. In *CVPR*, 2015. 5

[30] Wei Liu, Jun Wang, Sanjiv Kumar, and Shih-Fu Chang. Hashing with graphs. In *ICML*, 2011. 5

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5

[32] Alexandre Sablayrolles, Matthijs Douze, Nicolas Usunier, and Hervé Jégou. How should we evaluate supervised hashing? In *ICASSP*, 2017. 5

[33] 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. 2

[34] 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. 2

[35] 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. 2

[36] Paul Hongsuck Seo, Geeho Kim, and Bohyung Han. Combinatorial inference against label noise. In *NeurIPS*, 2019. 1, 2

[37] Paul Hongsuck Seo, Tobias Weyand, Jack Sim, and Bohyung Han. Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps. In *ECCV*, 2018. 2

[38] Fumin Shen, Chunhua Shen, Wei Liu, and Heng Tao Shen. Supervised discrete hashing. In *CVPR*, 2015. 5

[39] 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. 2

[40] 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. 2

[41] Rafael Vareto, Samira Silva, Filipe Costa, and William Robson Schwartz. Towards open-set face recognition using hashing functions. In *IJCB*, 2017. 2

[42] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *CVPR*, 2022. 2, 7, 8

[43] Guan'an Wang, Qinghao Hu, Jian Cheng, and Zengguang Hou. Semi-supervised generative adversarial hashing for image retrieval. In *ECCV*, 2018. 5

[44] 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. 5

[45] Rongkai Xia, Yan Pan, Hanjiang Lai, Cong Liu, and Shuicheng Yan. Supervised hashing for image retrieval via image representation learning. In *AAAI*, 2014. 5

[46] Hong Xuan, Richard Souvenir, and Robert Pless. Deep randomized ensembles for metric learning. In *ECCV*, 2018. 2

[47] Xinyu Yan, Lijun Zhang, and Wu-Jun Li. Semi-supervised deep hashing with a bipartite graph. In *IJCAI*, 2017. 2

[48] Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Naemura. Classification-reconstruction learning for open-set recognition. In *CVPR*, 2019. 2

[49] Tan Yu, Junsong Yuan, Chen Fang, and Hailin Jin. Product quantization network for fast image retrieval. In *ECCV*, 2018. 2, 3

[50] Andrew Zhai and Hao-Yu Wu. Classification is a strong baseline for deep metric learning. In *BMVC*, 2019. 3

[51] 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. 2, 5

[52] Bingchen Zhao and Kai Han. Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. In *NeurIPS*, 2021. 7, 8

[53] Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. Neighborhood contrastive learning for novel class discovery. In *CVPR*, 2021. 7, 8