Divide and Conquer: Compositional Experts for Generalized Novel Class Discovery

Muli Yang, Yuehua Zhu, Jiaping Yu, Aming Wu*, and Cheng Deng*
School of Electronic Engineering, Xidian University, Xi'an 710071, China
{mlyang, yuehuazhu, jpyu}@stu.xidian.edu.cn, amwu@xidian.edu.cn, chdeng@mail.xidian.edu.cn

Abstract

In response to the explosively-increasing requirement of annotated data, Novel Class Discovery (NCD) has emerged as a promising alternative to automatically recognize unknown classes without any annotation. To this end, a model makes use of a base set to learn basic semantic discriminability that can be transferred to recognize novel classes. Most existing works handle the base and novel sets using separate objectives within a two-stage training paradigm. Despite showing competitive performance on novel classes, they fail to generalize to recognizing samples from both base and novel sets. In this paper, we focus on this generalized setting of NCD (GNCD), and propose to divide and conquer it with two groups of Compositional Experts (ComEx). Each group of experts is designed to characterize the whole dataset in a comprehensive yet complementary fashion. With their union, we can solve GNCD in an efficient end-to-end manner. We further look into the drawback in current NCD methods, and propose to strengthen ComEx with global-to-local and local-to-local regularization. ComEx[1] *is evaluated on four popular benchmarks, showing clear superiority towards the goal of GNCD.*

1. Introduction

A key to the success of deep learning is huge amounts of curated data with elaborate annotations [12, 25, 31]. Despite being expensive and cumbersome to collect, the annotated data plays an indispensable role since deep models are notoriously known to be data-hungry. In this regard, *Semi-Supervised Learning (SSL)* [4, 5, 39] sheds some light on the dilemma. Requiring only a small amount of annotations, SSL addresses unannotated data using *pseudo-labeling* or *consistency regularization*, yet limited to known classes from existing annotations. To recognize unknown classes never seen in training, *Zero-Shot Learning (ZSL)* [9, 28, 37] resorts to extra annotations to learn transferable attributes among known and unknown classes, which in turn exacerbates the need for annotations.

We consider a new task setting that naturally addresses the limitations of both SSL and ZSL, namely *Novel Class Discovery (NCD)* [15, 16]. As shown in Fig. 1a, NCD assumes two sets of samples with disjoint classes — a *base set* containing *labeled* samples of *base* classes, and a *novel set* containing *unlabeled* samples of *novel* classes. The goal of NCD, besides correctly classifying base[2] samples, is to recognize novel classes out of unlabeled samples using no extra knowledge. In practice, the base and novel sets are often subsets of a same dataset, and thus share similar visual knowledge that can be transferred to discover novel classes.

In view of this, most existing NCD methods [15, 16, 24, 46–48] adopt a two-stage paradigm: a supervised training stage on the base set to learn basic semantic discriminability, and a fine-tuning stage on the novel set to discover novel classes by clustering unlabeled samples. To compensate the lack of supervised information, the two aforementioned SSL techniques are often employed to strengthen clustering performance — using pseudo labels estimated in the novel set as clustering targets, and enforcing consistency between different transformations of a same input. Despite being competitive on discovering novel classes, these methods inevitably suffer from performance degradation on the base set due to separate objectives for base and novel classes. This can be problematic for deployment on a system handling data from both sets. A most recent work [13] identifies this problem and proposes a unified objective for NCD. Although yielding promising results on either base or novel set, this method struggles to generalize to the union of the two sets, *i.e.*, testing on both base and novel samples, showing it is still not yet ready for real-world deployment.

In this paper, we focus on the generalized setting of NCD (GNCD), aiming at designing a unified model that works well on both base and novel sets, especially on their union. This is challenging due to the uneven properties of

*Corresponding authors.

[1] Code: https://github.com/muliyangm/ComEx.

[2] We interchangeably use *base / labeled*, and *novel / unlabeled*.

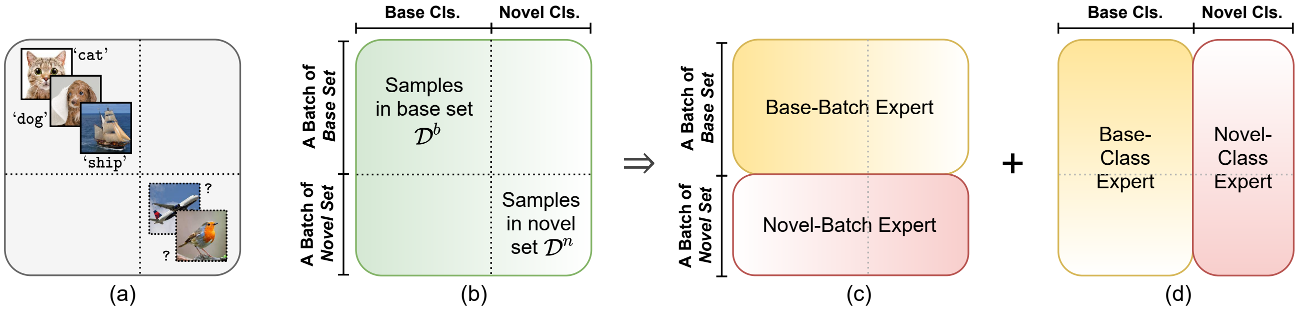

Figure 1. **(a)** A training batch for Novel Class Discovery (NCD). **(b)** A *batch-class* view of NCD. We can decompose a batch of training samples with *batch-wise* and *class-wise* perspectives. "Cls." is short for "Classes". **(c)** Batch-wise experts, dealing with respective sub-batches, yet aware of both base and novel classes. **(d)** Class-wise experts, handling a whole batch of samples, but with class-wise expertise.

the two sets. As shown in Fig. 1b, we seek to leverage the compositional nature of base and novel sets by viewing GNCD within a *batch-class* perspective. By respectively decomposing base and novel sets according to batches and classes, we propose to divide and conquer GNCD with two groups of *Compositional Experts (ComEx)*. As illustrated in Figs. 1c and 1d, each group of experts characterizes the whole dataset in a comprehensive yet complementary way — with batch-wise experts (Fig. 1c) capturing *separability* between base and novel classes, and class-wise experts (Fig. 1d) modeling *discriminability* within each set of classes. With their union, we can achieve our goal of GNCD in an end-to-end manner (see Sec. 3.1). Inside the experts, we regard the weights of each clustering head as a series of global cluster centers, such that our experts can be powered by sophisticated pseudo-labeling technique [1, 8], which actually induces a *global-to-local* alignment between global cluster centers and local training samples. We argue that this pure global-to-local formulation can be vulnerable to local changes (*e.g*., color, background), resulting in unfavorable clustering performance. To this end, we introduce local consistency into pseudo labels by aggregating neighborhood information according to soft similarities, *i.e*., a *local-to-local* aggregation (see Sec. 3.2), which offers clear performance boost on novel classes.

To sum up, our contributions are threefold:

- Our work is among the first attempts that focus on the generalized setting of NCD, and proposes to divide and conquer it using compositional experts that characterize the data in a unique yet complementary manner;
- We interpret pseudo-labeling as a global-to-local alignment between cluster centers and training samples, and propose to strengthen pseudo labels with local-to-local aggregation among neighborhood samples;
- We show in extensive experiments the superiority of our proposed method against several state of the arts, and push the limits of NCD into the challenging generalized setting that is of greater practical significance.

2. Related Work

Novel Class Discovery. The aim of Novel Class Discovery (NCD) [15, 16] is to utilize the knowledge of base classes to discover novel classes by forming clusters on novel samples. Typically, NCD takes both labeled (base classes) and unlabeled (novel classes) samples, performing supervised classification and unsupervised clustering on them, respectively. NCD is in concept related to Zero-Shot Learning (ZSL), Semi-Supervised Learning (SSL), and unsupervised clustering, but is also significantly different from them. In particular, ZSL [3, 9, 28, 30, 37, 40, 43] aims to recognize novel classes never seen in training, relying on auxiliary semantic attributes to infer class relations, which are absent in NCD. SSL [2, 4, 5, 29, 33, 39] also follows a labeled-and-unlabeled training paradigm, with the assumption that all unlabeled samples are from the classes of labeled samples, while NCD assumes no class-overlap between labeled and unlabeled samples. Compared to unsupervised clustering [1, 7, 22] that incorporates no extra supervision, NCD makes use of a base set to exploit semantic guidance for better clustering performance on the novel set.

To address NCD, early works [20, 21] mainly focus on estimating pairwise similarities among novel samples using the prediction ability learned on the base set. These pairwise similarities are then used to train a clustering model for recognizing novel classes. In a later work [16], NCD training is standardized within a two-stage procedure: 1) a fully-supervised training stage on base samples, and 2) a fine-tuning stage on novel samples. This two-stage training procedure is adopted by most of the follow-up works [13, 15, 24, 46–48]. Specifically, Han *et al*. [15] further amplified the first training stage using a self-supervised pretext task. Moreover, they resorted to ranking statistics for measuring pairwise similarities between novel samples;

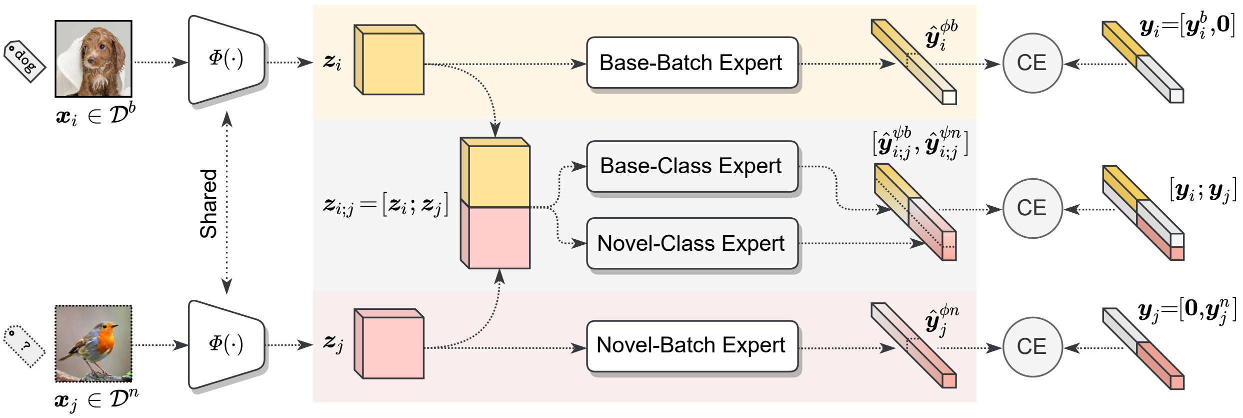
dog
$\boldsymbol{x}_i \in \mathcal{D}^b$
$\Phi(\cdot)$
Shared
?
$\boldsymbol{x}_j \in \mathcal{D}^n$
$\Phi(\cdot)$
$\boldsymbol{z}_i$
$\boldsymbol{z}_{i;j} = [\boldsymbol{z}_i; \boldsymbol{z}_j]$
$\boldsymbol{z}_j$
Base-Batch Expert
Base-Class Expert
Novel-Class Expert
Novel-Batch Expert
$\hat{\boldsymbol{y}}_i^{\phi b}$
$[\hat{\boldsymbol{y}}_{i;j}^{\psi b}, \hat{\boldsymbol{y}}_{i;j}^{\psi n}]$
$\hat{\boldsymbol{y}}_j^{\phi n}$
CE
CE
CE
$\boldsymbol{y}_i = [\boldsymbol{y}_i^b, \mathbf{0}]$
$[\boldsymbol{y}_i; \boldsymbol{y}_j]$
$\boldsymbol{y}_j = [\mathbf{0}, \boldsymbol{y}_j^n]$

Figure 2. Schematic of our proposed ComEx. We sample a batch of images from both base and novel sets for training; in this figure we only show one image in each set for brevity. The images are passed into a shared encoder, and then fed into the two groups of experts to obtain predictive outputs. Note that both groups of experts share a same target for each input. "CE" is short for the cross-entropy loss.

the same metric is also adopted in [24, 46]. Zhong *et al.* proposed to augment the pairwise similarities among novel samples using neighborhood contrastive learning [47], or generated virtual samples [48] with MixUp [45]. Benefiting from the optimal transport style pseudo-label assignment [1, 8], Fini *et al.* [13] managed to unify the learning of both base and novel classes using a single cross-entropy loss. They also proposed a "task-agnostic" testing protocol that measures the generalization ability when not knowing which class subset a testing sample is from, forming the idea of Generalized NCD (GNCD), which is of greater practical significance, yet far from being solved. In this paper, we follow [13] to tackle the more challenging GNCD problem, with our proposed compositional experts simultaneously capturing separability between base and novel classes, and discriminability within them.

Unsupervised Clustering. Our work is closely related to unsupervised clustering [14, 38, 41, 44], which aims to partition an unlabeled dataset into different clusters given no semantic supervision. To this end, a common choice is neighborhood aggregation [11, 22, 35, 49], which is based on the assumption that data points within a neighborhood feature space likely share a same semantic label. By enforcing neighborhood consistency and average entropy maximization, one can achieve clustering as well as avoiding cluster collapse. Another line of works resort to pseudo-labeling for simultaneously learning feature representations and clustering. Caron *et al.* [7] showed that k-means steps can naturally produce pseudo labels for unsupervised representation learning. Asano *et al.* [1] advanced pseudo-labeling by considering it as an optimal transport problem, which can be effectively solved using the Sinkhorn-Knopp algorithm [10]. This formulation was further developed into a powerful self-supervised learning method [8] by contrasting pseudo-label assignments between different augmentations of an image. In this paper, we also resort to this pseudo-labeling strategy [1, 8] for uniformly training on both base and novel classes. By further looking into the pseudo-labeling behavior, we propose to strengthen its original global-to-local alignment with local-to-local aggregation based on similarities between neighborhood samples.

3. Approach

We present in this section our proposed ComEx. We start with the problem definition of NCD, followed by the overall look of our solution, and its detailed formulation.

Problem Definition. In NCD, training data are divided into a base set and a novel set. The base set $\mathcal{D}^b = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N^b}$ contains samples $\boldsymbol{x}_i$ associated with corresponding labels y_i from C^b classes. The novel set $\mathcal{D}^n = \{\boldsymbol{x}_j\}_{j=N^b+1}^{N^b+N^n}$ contains unlabeled samples $\boldsymbol{x}_j$ from C^n novel classes, in which C^n is known a prior. Classes in base and novel sets are non-overlapped. The aim of NCD is to discover unknown classes in the novel set, *i.e.*, by partitioning $\mathcal{D}^n$ into C^n clusters using the knowledge learned from $\mathcal{D}^b$. Other than that, in this paper we focus on the more challenging GNCD problem, in which a testing sample comes from both sets, yet we do not know in advance which set it is from.

Overall Framework. Our goal is to learn a mapping from the image space $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^{N^b+N^n}$ to the joint label space $\mathcal{Y} = \{l\}_{l=1}^{C^b+C^n}$. At each training step, we sample a batch of images from both $\mathcal{D}^b$ and $\mathcal{D}^n$. As shown in Fig. 2, for each input $\boldsymbol{x}$, we pass it to a shared image encoder $\Phi(\cdot)$ to extract its visual feature, *i.e.*, $\boldsymbol{z} = \Phi(\boldsymbol{x})$. The visual feature is then fed into two groups of experts (batch- and class-wise) for deriving the predictive outputs, which are then trained with the cross-entropy loss. For samples from $\mathcal{D}^b$, we directly use their ground-truth labels as training targets; for those from $\mathcal{D}^n$, we use pseudo-labeling to generate the targets, which will be introduced in Sec. 3.2. Note that all experts are composed of shallow-layer MLPs, only requiring a small amount of extra computational resources. For inference, we simply combine the outputs of these experts.

3.1. Compositional Experts

As shown in Figs. 1c and 1d, we design the compositional experts in batch-wise and class-wise, respectively. Our design follows the "*known unknown*" [42] principle, *i.e.*, each expert should know when it does not know, which facilitates their cooperation by eliminating unwanted conflicts. This principle enables each group of experts accessing to the whole dataset, allowing them to capture comprehensive yet complementary information.

Batch-wise Experts. Each batch-wise expert deals with its corresponding sub-batch of samples. For an input $\boldsymbol{x}_i \in \mathcal{D}^b$, we directly feed its extracted feature $\boldsymbol{z}_i$ into the *base-batch expert* — parameterized by a linear classifier $\phi^b(\cdot)$ with $C^b + C^n$ output neurons — to obtain the predictive output, *i.e.*, $\hat{\boldsymbol{y}}_i^{\phi b} = \phi^b(\boldsymbol{z}_i)$. Likewise, taking as input the visual feature $\boldsymbol{z}_j$ of $\boldsymbol{x}_j \in \mathcal{D}^n$, the *novel-batch expert* first projects $\boldsymbol{z}_j$ into a low-dimensional representation $\boldsymbol{z}_j^{\phi}$ using a MLP, and then passes it to another linear classifier $\phi^n(\cdot)$ (also with $C^b + C^n$ output neurons) to get the output, *i.e.*, $\hat{\boldsymbol{y}}_j^{\phi n} = \phi^n(\boldsymbol{z}_j^{\phi}), \boldsymbol{z}_j^{\phi} = \phi'^n(\boldsymbol{z}_j)$. Note that both features $\boldsymbol{z}_i, \boldsymbol{z}_j^{\phi}$ and the weights of linear classifiers $\phi^b(\cdot), \phi^n(\cdot)$ are ℓ_2-normalized when computing the predictive outputs.

We use a standard cross-entropy loss to train the two experts. After obtained the batch-wise expert output $\hat{\boldsymbol{y}}^{\phi} \in \mathbb{R}^{C^b+C^n}$ of an input $\boldsymbol{x}$, the cross-entropy loss is defined as

$$\mathcal{L}_{ce}(\hat{\boldsymbol{y}}^{\phi}, \boldsymbol{y}) = -\boldsymbol{y} \log \sigma(\hat{\boldsymbol{y}}^{\phi}/\tau)\,, \tag{1}$$

where $\boldsymbol{y} \in \mathbb{R}^{C^b+C^n}$ is a one-hot label with its first C^b and last C^n elements corresponding to base and novel classes, respectively; $\sigma(\cdot)$ is a softmax function, and τ is the temperature parameter [19]. This loss naturally holds for base samples; for novel samples without ground-truth labels, we generate pseudo labels as their training targets (see Sec. 3.2).

By designing, each batch-wise expert handles disjoint part of the dataset. In other words, for base-batch expert, the last C^n elements of its output $\hat{\boldsymbol{y}}^{\phi b}$ should always being suppressed — to prevent erroneously predicting a base class into novel classes; the same with the first C^b elements of the novel-batch expert output $\hat{\boldsymbol{y}}^{\phi n}$. Although the cross-entropy loss in Eq. (1) implicitly achieves this goal by setting these targets to all zeros, it makes no distinction between base and novel classes. Similar to [6], we introduce extra regularization to explicitly suppress the non-target outputs:

$$\mathcal{L}_{reg}(\hat{\boldsymbol{y}}^{\phi}) = \left(\sum\nolimits_{c \in \widetilde{\mathcal{Y}}} (\hat{y}_c^{\phi})^2\right)^{\frac{1}{2}}\,, \tag{2}$$

where $\widetilde{\mathcal{Y}} = \{1, 2, \cdots, C^b\}$ for outputs from the novel-batch expert, $\widetilde{\mathcal{Y}} = \{C^b{+}1, C^b{+}2, \cdots, C^b{+}C^n\}$ for outputs from the base-batch expert, and $\hat{y}_c^{\phi}$ is the c-th element of $\hat{\boldsymbol{y}}^{\phi}$.

Class-wise Experts. As shown in Fig. 2, the class-wise experts are applied to whole batch of training samples. Similar to the base-batch experts, the *base-class expert* is also implemented as a linear classifier $\psi^b(\cdot)$, but with C^b output neurons; its output is written as $\hat{\boldsymbol{y}}^{\psi b} = \psi^b(\boldsymbol{z})$, where $\hat{\boldsymbol{y}}^{\psi b} \in \mathbb{R}^{C^b}$. Likewise, the *novel-class expert* also first projects $\boldsymbol{z}$ to a low-dimensional $\boldsymbol{z}^{\psi}$, and then pass it to another linear classifier $\psi^n(\cdot)$ (with C^n output neurons) to get the output, *i.e.*, $\hat{\boldsymbol{y}}^{\psi n} = \psi^n(\boldsymbol{z}^{\psi}), \boldsymbol{z}^{\psi} = \psi'^n(\boldsymbol{z})$, where $\hat{\boldsymbol{y}}^{\psi n} \in \mathbb{R}^{C^n}$. Visual features and linear classifiers are also ℓ_2-normalized, similar to the batch-wise experts.

A straightforward way to train the two class-wise experts is using separate cross-entropy losses. However, this breaks our "known unknown" principle such that the learned class prototypes of the two experts inevitably mingle with each other in the output space, which gives rise to unwanted conflicts when aggregating them for inference. This is also empirically verified in [13]. In view of this, we instead use a single loss that regularizes the joint class space:

$$\mathcal{L}_{ce}\left([\hat{\boldsymbol{y}}^{\psi b}, \hat{\boldsymbol{y}}^{\psi n}], \boldsymbol{y}\right)\,, \tag{3}$$

in which both variables of the cross-entropy loss are in $\mathbb{R}^{C^b+C^n}$, and we leave the construction of target $\boldsymbol{y}$ to Sec. 3.2. We have also tried to apply extra regularization on class-wise experts to explicitly suppress the non-target outputs similar to Eq. (2), *e.g.*, base-class expert outputs of novel samples. This turned out to be harmful in experiments, and an explanation can be that rigidly zeroing outputs of a certain amount of inputs results in oversensitiveness to data.

3.2. Training Targets

Now we introduce how training targets are constructed, *i.e.*, $\boldsymbol{y}$ in Eqs. (1) and (3). For a given input $\boldsymbol{x}$, both batch- and class-wise experts share the same target $\boldsymbol{y} \in \mathbb{R}^{C^b+C^n}$. As shown in Fig. 2, when $\boldsymbol{x}_i$ is from the base set $\mathcal{D}^b$, with its ground-truth label as y_i, the target $\boldsymbol{y}_i$ is constructed as its one-hot representation, *i.e.*, the y_i-th element of $\boldsymbol{y}_i$ being 1 and the rest $C^b{+}C^n{-}1$ elements being all zeros, also written as $\boldsymbol{y}_i = [\boldsymbol{y}_i^b, \boldsymbol{0}]$. When $\boldsymbol{x}_j$ is from the novel set $\mathcal{D}^n$, due to the lack of ground-truth label, we construct the target $\boldsymbol{y}_j$ by setting its first C^b elements to all zeros, and the rest C^n ones using pseudo-labeling, *i.e.*, $\boldsymbol{y}_j = [\boldsymbol{0}, \boldsymbol{y}_j^n]$.

Pseudo-Labeling. We follow [1, 8, 13] to generate pseudo labels for samples in $\mathcal{D}^n$. By regarding the weight parameters of each linear classifier $\phi^n(\cdot), \psi^n(\cdot)$ in novel-batch and novel-class experts as a series of *global cluster centers* (or class prototypes), our goal is to equally partition a batch of samples to these cluster centers, as well as maximizing the similarities between visual features and cluster centers.

Taking the novel-batch expert as an example, we assemble the visual features of a batch of novel samples into a matrix $\mathbf{Z}=[\boldsymbol{z}_1^{\phi};\dots;\boldsymbol{z}_{B^n}^{\phi}]\in\mathbb{R}^{B^n\times d}$, where B^n is the number of novel samples in a batch and d is the feature dimension. The weight parameters of the linear classifier $\phi^n(\cdot)$, corresponding to C^n novel classes, are taken as the cluster centers $\mathbf{W}=[\boldsymbol{w}_1;\dots;\boldsymbol{w}_{C^n}]\in\mathbb{R}^{C^n\times d}$. Our goal can be achieved by solving

$$\max_{\mathbf{Y}\in\mathcal{P}}\operatorname{tr}(\mathbf{Y}\mathbf{W}\mathbf{Z}^{\top})+\epsilon H(\mathbf{Y})\,, \tag{4}$$

where $\mathbf{Y}=[\boldsymbol{y}_1^{\phi n};\dots;\boldsymbol{y}_{B^n}^{\phi n}]\in\mathbb{R}_+^{B^n\times C^n}$ should equally map B^n novel samples to C^n cluster centers while achieving maximum similarities, $\mathcal{P}$ is a search space ensuring each cluster center to be selected at least $\frac{B^n}{C^n}$ times on average, $H(\cdot)$ is the entropy function used to control the smoothness of $\mathbf{Y}$, and ϵ is a trade-off parameter. We refer the reader to [1,8] for more details. The solution $\mathbf{Y}$ to Eq. (4), which can be calculated using the iterative Sinkhorn-Knopp algorithm [10], is taken as soft pseudo labels for the B^n novel samples. The target $\boldsymbol{y}^{\psi n}$ for the novel-class expert can be calculated the same way.

Local Aggregation. The above pseudo-labeling method, powered by optimal transport, can be seeing as a *global-to-local* alignment between cluster centers and training samples. This method, however, lacks consideration of local relations among different data samples, which may result in vulnerability to local changes, such as colors and backgrounds. Inspired by neighborhood-based clustering [22,35,49] and semi-supervised learning [2,29] methods that aggregate k-nearest neighbors (k-NNs), we propose a *local-to-local* aggregation strategy to strengthen the consistency among neighborhood samples.

To this end, we maintain an offline first-in-first-out queue [17] to allow cross-batch aggregation in training. As shown in Fig. 3, the queue stores features and pseudo labels of most recent N^q novel samples into a dictionary $\mathcal{Q}=\{\boldsymbol{z}_k^q\colon\boldsymbol{y}_k^q\}_{k=1}^{N^q}$, where $\boldsymbol{z}^q$ is the mean of $\boldsymbol{z}^{\phi}$ and $\boldsymbol{z}^{\psi}$, and $\boldsymbol{y}^q$ is the mean of $\boldsymbol{y}^{\phi n}$ and $\boldsymbol{y}^{\psi n}$. For each novel sample $\boldsymbol{x}$ in training, we calculate softmax-normalized similarities between its visual feature $\bar{\boldsymbol{z}}$ (mean of $\boldsymbol{z}^{\phi}$ and $\boldsymbol{z}^{\psi}$) and each queue feature $\boldsymbol{z}^q$:

$$s_k=\frac{\exp(\bar{\boldsymbol{z}}\cdot\boldsymbol{z}_k^q/\tau)}{\sum_{k=1}^{N^q}\exp(\bar{\boldsymbol{z}}\cdot\boldsymbol{z}_k^q/\tau)}\,, \tag{5}$$

where both $\bar{\boldsymbol{z}}$ and $\boldsymbol{z}^q$ are ℓ_2-normalized. Instead of selecting k-NNs, we use a softer manner for aggregation:

$$\boldsymbol{y}^n=\alpha\frac{\boldsymbol{y}^{\phi n}+\boldsymbol{y}^{\psi n}}{2}+(1-\alpha)\sum\nolimits_{k=1}^{N^q}s_k\boldsymbol{y}_k^q\,, \tag{6}$$

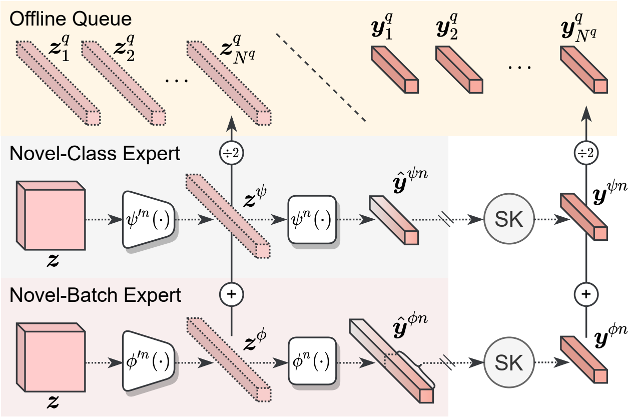

Figure 3. Construction of queue $\mathcal{Q}$. Given visual feature $\boldsymbol{z}$ of an input image, the novel-batch/class expert first projects it into a low-dim feature, and then outputs predictive logits, which are used to generate pseudo labels. Low-dim features and pseudo labels of most recent N^q samples are preserved in the queue as the mean of two experts. "SK" is short for the Sinkhorn-Knopp algorithm, and the gradient stops when calculating pseudo labels in training.

which integrates pseudo labels of neighborhood samples in the queue into the current training sample according to similarities, and $\alpha\in[0,1]$ is a hyperparameter controlling the intensity of local aggregation.

Our final training target $\boldsymbol{y}^n$ explicitly encourages local consistency among novel samples while maintaining a global alignment between samples and cluster centers. Eq. (6) can also be interpreted as a "query-key-value" attention [36] with a shortcut, which allows a flexible information flow among neighborhood samples and cluster centers.

3.3. Overall Objective

All the four experts are jointly trained in an end-to-end manner. The overall training loss can be written as

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{y})=\mathcal{L}_{ce}(\hat{\boldsymbol{y}}^{\phi},\boldsymbol{y})+\mathcal{L}_{ce}(\hat{\boldsymbol{y}}^{\psi},\boldsymbol{y})+\mathcal{L}_{reg}(\hat{\boldsymbol{y}}^{\phi})\,, \tag{7}$$

where $\hat{\boldsymbol{y}}^{\phi}=\hat{\boldsymbol{y}}^{\phi b},\boldsymbol{y}=[\boldsymbol{y}^b,\mathbf{0}]$ for $\boldsymbol{x}\in\mathcal{D}^b$; when $\boldsymbol{x}\in\mathcal{D}^n$, we have $\hat{\boldsymbol{y}}^{\phi}=\hat{\boldsymbol{y}}^{\phi n},\boldsymbol{y}=[\mathbf{0},\boldsymbol{y}^n]$; for both cases, $\hat{\boldsymbol{y}}^{\psi}=[\hat{\boldsymbol{y}}^{\psi b},\hat{\boldsymbol{y}}^{\psi n}]$. At each training step, we follow recent works [13, 47, 48] to first generate two views $\boldsymbol{v}_1,\boldsymbol{v}_2$ for each input image $\boldsymbol{x}$ using data augmentations. Accordingly, our model outputs two predictions associated with two targets $\boldsymbol{y}_1,\boldsymbol{y}_2$. To encourage consistency across views, we adopt the swapped prediction strategy [8,13] by training with $\mathcal{L}(\boldsymbol{v}_1,\boldsymbol{y}_2)+\mathcal{L}(\boldsymbol{v}_2,\boldsymbol{y}_1)$ for each input. For inference, we simply combine the outputs of four experts:

$$\hat{\boldsymbol{y}}=\hat{\boldsymbol{y}}^{\phi b}+\hat{\boldsymbol{y}}^{\phi n}+[\hat{\boldsymbol{y}}^{\psi b},\hat{\boldsymbol{y}}^{\psi n}]\,, \tag{8}$$

in which the first two terms correspond to the contribution of batch-wise experts, and the last term corresponds to that of the class-wise ones.

Dataset →					CIFAR10						CIFAR100-50					
Evaluation protocol →					Task-aware			Task-agnostic			Task-aware			Task-agnostic		
#	$\phi(\cdot)$	$\mathcal{L}_{reg}$	$\psi(\cdot)$	$\mathcal{Q}$	Base	Nov.	All	Base	Nov.	All	Base	Nov.	All	Base	Nov.	All
1	✓				96.6	91.2	93.9	90.0	89.8	89.9	79.6	49.9	64.8	74.6	49.0	61.8
2	✓	✓			96.2	92.6	94.4	93.3	89.7	91.5	79.8	49.7	64.8	75.1	48.8	62.0
3	✓			✓	96.4	93.1	94.8	90.5	91.0	90.8	79.9	53.1	66.5	74.8	52.3	63.6
4			✓		96.6	91.1	93.9	93.0	88.9	91.0	79.4	50.2	64.8	71.2	48.2	59.7
5			✓	✓	96.3	93.0	94.7	93.0	90.2	91.6	79.9	52.7	66.3	71.2	51.4	61.3
6	✓		✓		96.6	91.1	93.9	94.8	90.6	92.7	79.6	50.1	64.9	74.5	49.9	62.2
7	✓	✓	✓		96.3	92.7	94.5	94.4	91.8	93.1	79.8	50.3	65.1	75.2	49.9	62.6
8	✓		✓	✓	96.3	93.1	94.7	94.8	92.4	93.6	80.1	53.7	66.9	75.0	53.3	64.2
9	✓	✓	✓	✓	**96.7**	**93.2**	**95.0**	**95.0**	**92.6**	**93.8**	**80.2**	**54.2**	**67.2**	**75.3**	**53.5**	**64.4**

Table 1. Ablation study of ComEx. Results are reported in classification/clustering accuracy (%) on base/novel set. We use task-aware ***(a2, a3)*** and task-agnostic ***(b1)*** protocols for evaluation. **Best** results are highlighted in each column. "Nov." is short for "Novel".

Subset →	Base		Novel	
Dataset ↓	Images	Classes	Images	Classes
CIFAR10	25K	5	25K	5
CIFAR100-20	40K	80	10K	20
CIFAR100-50	25K	50	25K	50
ImageNet	1.25M	882	≈30K	30

Table 2. Datasets statistics in terms of base/novel subsets.

4. Experiments

Datasets. We evaluate our proposed ComEx on three widely-used NCD benchmark datasets, *i.e.*, CIFAR10 [26], CIFAR100 [26], and ImageNet [12]. As shown in Tab. 2, each dataset is split into two subsets, namely a base set and a novel set. The base set contains labeled images of base classes, while the novel set contains unlabeled images of novel classes (non-overlap with the base classes). We conduct experiments on four different data splits of the three datasets as shown in Tab. 2, in which *CIFAR100-50* was newly introduced in [13] for a more challenging evaluation on novel classes, and the rest three splits are widely adopted in former works. We assume the number of novel classes known a priori during training following recent works [13, 15, 24, 46–48]. This assumption facilitates our focus on designing a unified model for both base and novel classes. Note that if necessary, we can also opt for the off-the-shelf method [16] to estimate novel class numbers. Other than the class numbers, no supervision is used for the novel set during training.

Evaluation Metrics. Following [13], we evaluate our proposed method using two protocols: ***(a)*** *task-aware*, in which the subset (*i.e.*, base or novel set) of each testing sample is known in advance; and ***(b)*** *task-agnostic*, in which the subset information is unknown. In experiments, the base/novel set is further divided into training and testing splits. For the task-aware protocol, we can evaluate the performance on ***(a1)*** *training split of the novel set*, ***(a2)*** *testing split of the novel set*, and ***(a3)*** *testing split of the base set*. In contrast, for the task-agnostic protocol, the evaluation takes place in ***(b1)*** *both testing splits of the base and novel sets*.

Typically, most recent works [24, 47, 48] only evaluate their methods on *(a1)* with a transductive learning fashion. In contrast, we also involve evaluations on *(a2)*, *(a3)*, and *(b1)* to meet the aim of Generalized NCD as we discussed in Secs. 1 and 2. In particular, to evaluate the classification performance on the base set, we simply measure the classification accuracy. For evaluating the clustering performance on the novel set, we adopt clustering accuracy following recent works, which is defined as

$$\mathrm{Acc} = \max_{g \in \mathcal{P}(C^n)} \frac{1}{N} \sum\nolimits_{j=1}^{N} \mathbb{1}\{y_j = g(\hat{y}_j)\}\,, \tag{9}$$

where y_j and $\hat{y}_j$ are respectively ground-truth label and cluster assignment prediction for each novel sample $\boldsymbol{x}_j$; N is the total number of novel samples for testing; $\mathcal{P}(C^n)$ denotes the set of all possible permutations of C^n elements, and g is an arbitrary permutation. The optimal permutation g^* can be obtained using the Hungarian algorithm [27].

Implementation Details. We use a ResNet-18 [18] as the image encoder for fair comparisons with the existing works. The base-batch and base-class experts, *i.e.*, $\phi^b(\cdot)$ and $\psi^b(\cdot)$, are both linear classifiers with $C^b + C^n$ and C^n output neurons, respectively. The novel-batch and novel-class experts are both composed of an MLP that maps 512-dimensional visual features to 256-dimensional ones with 2048 hidden units, followed by a linear classifier that outputs $C^b + C^n$ or C^n dimensional logits, respectively. We follow [7, 13, 23] to use over-clustering and multi-head clustering strategies for better clustering performance. For image augmentations

Dataset →	CIFAR10			CIFAR100-20			CIFAR100-50		
Method ↓	Base	Novel	All	Base	Novel	All	Base	Novel	All
KCL [20]	79.4	60.1	69.8	23.4	29.4	24.6	–	–	–
MCL [21]	81.4	64.8	73.1	18.2	18.0	18.2	–	–	–
DTC [16]	58.7	78.6	68.7	47.6	49.1	47.9	30.2	34.7	32.5
RS+ [15]	90.6	88.8	89.7	71.2	56.8	68.3	69.7	40.9	55.3
UNO [13]	93.6†	89.9†	91.8†	73.2	73.1	73.2	71.5	50.7	61.1
ComEx (Ours)	**95.0**+1.4	**92.6**+2.7	**93.8**+2.0	**75.2**+2.0	**77.3**+4.2	**75.6**+2.4	**75.3**+3.8	**53.5**+2.8	**64.4**+3.3

Table 3. Comparison with state of the arts. Results are reported in classification/clustering accuracy (%) on testing split of base/novel set using task-agnostic ***(b1)*** evaluation protocol. **Best** and second-best results are highlighted in each column. †Our reproduced result.

we use moderate random crop, flip, jittering, and grey-scale following [13]. Please see Appendix for more details.

4.1. Ablation Study

We ablate our proposed ComEx to evaluate the effectiveness of each proposed module, including the batch-wise experts (denoted by $\phi(\cdot)$ for brevity), the class-wise experts ($\psi(\cdot)$), the regularization $\mathcal{L}_{reg}$ in Eq. (2), and the queue $\mathcal{Q}$ for local aggregation defined in Eqs. (5) and (6). The ablation study is conducted on the testing split of both base and novel sets, using task-aware *(a2, a3)* and task-agnostic *(b1)* evaluation protocols. The results are summarized in Tab. 1. As can be seen, in general, each proposed module contributes to the full model. We present below detailed discussions with respective to two different aspects, namely the effect of different experts (as well as the regularization $\mathcal{L}_{reg}$) and the effect of the queue for local aggregation.

Effect of Experts. We evaluate the effect of batch-wise experts $\phi(\cdot)$ by disabling the class-wise ones $\psi(\cdot)$, and vice versa. Specifically, for each group of experts, we can inspect the performance of an individual expert (*e.g.*, the base- or novel-batch expert $\phi^b(\cdot), \phi^n(\cdot)$) using task-aware evaluations; their collaborative performance can be evaluated using the task-agnostic protocol. In Tab. 1, we report in #1–3 the performance of solely using batch-wise experts, in #4–5 the performance of only using class-wise experts, and in #6–9 the performance of using both groups of experts, in which #9 is our full model.

We can observe that each group of experts performs stably across datasets. Interestingly, compared to the class-wise experts, in #1–3 the batch-wise experts demonstrate much stronger performance with task-agnostic evaluation protocol on CIFAR100-50, which leads to surprisingly high accuracy over current state of the arts (see Tab. 3). This is explainable since CIFAR100-50 contains much more novel classes to be clustered, which results in ambiguous boundaries between base and novel classes, making task-agnostic evaluations much more difficult. The good performance of batch-wise experts is attributed to our all-class-aware design, which explicitly induces separability between the two sets of classes — by guiding each batch-wise expert to learn from what it knows, and to reject what it does not know. With the regularization $\mathcal{L}_{reg}$ on non-target outputs we can observe further benefit towards this goal.

Method ↓	CF10	CF100-20	CF100-50	ImgNet
k-means [32]	72.5±0.0	56.3±1.7	28.3±0.7	71.9
KCL [20]	72.3±0.2	42.1±1.8	–	73.8
MCL [21]	70.9±0.1	21.5±2.3	–	74.4
DTC [16]	88.7±0.3	67.3±1.2	35.9±1.0	78.3
RS [15]	90.4±0.5	73.2±2.1	39.2±2.3	82.5
RS+ [15]	91.7±0.9	75.2±4.2	44.1±3.7	82.5
OpenMix [48]	**95.3**	–	–	85.7
DualRank [46]	91.6±0.6	75.3±2.3	–	88.9
Joint [24]	93.4±0.6	76.4±2.8	–	86.7
NCL [47]	93.4±0.5	**86.6±0.4**	–	90.7
UNO [13]	92.6±0.5†	85.0±0.6	52.9±1.4	90.6
ComEx (Ours)	93.6±0.3	85.7±0.7	**53.4±1.3**	**90.9**

Table 4. Comparison with state of the arts. Results are reported in clustering accuracy (%) on training split of the novel set using task-aware ***(a1)*** evaluation protocol. **Best** and second-best results are highlighted in each column. “CF” is short for “CIFAR”, and “ImgNet” is short for “ImageNet”. †Our reproduced result.

In contrast, the class-wise experts have particular expertise in the two sets of classes, which allows them to concentrate on their own specialties, and thus perform favorably in task-aware evaluations. By combining the two groups of experts, we can see further benefits thanks to their complementary abilities. This is especially obvious with the most challenging task-agnostic protocol, which evaluates not only separability between base and novel classes, but also discriminability within them.

Effect of Local Aggregation. We further evaluate the effect of our local aggregation strategy defined in Eqs. (5) and (6). By definition, this strategy alters the training target of each sample by aggregating neighborhood information preserved in an offline queue $\mathcal{Q}$. We report in Tab. 1 the results of performance with/without the queue. We can observe clear performance boosts on novel classes in both task-aware and task-agnostic evaluations when adding the queue. This veri-

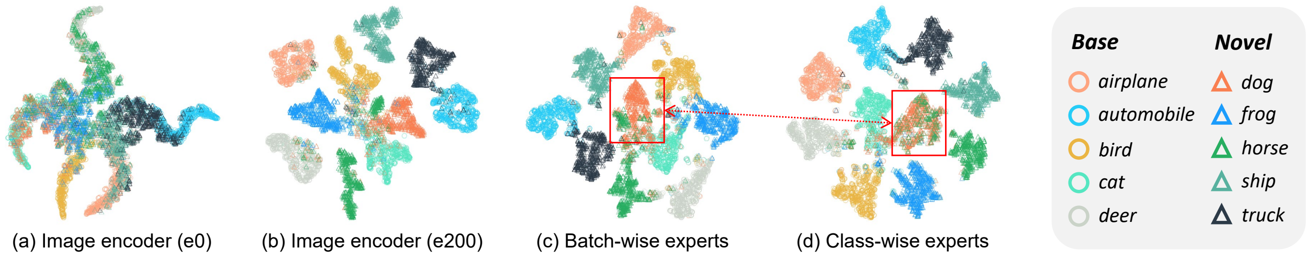

Figure 4. t-SNE [34] visualizations of ComEx on testing split of both base and novel sets of CIFAR10. **(a)** Initial feature space at training epoch 0. **(b)** Feature space after 200 training epochs. **(c)** Output space of batch-wise experts. **(d)** Output space of class-wise experts.

fies our assumption that incorporating local-to-local consistency provides further clustering benefits on top of global-to-local alignments. In experiments we observe that a moderate queue size of $N^q = 500$ is good enough for local aggregation. We also provide in Appendix the results using different queue size and different α in Eq. (6).

4.2. Comparison with State of the Arts

We compare our proposed ComEx with the current state-of-the-art NCD methods, *i.e.*, KCL [20], MCL [21], DTC [16], RS [15], RS+ [15] (RS with incremental classifier), OpenMix [48], DualRank [46], Joint [24], NCL [47], and UNO [13]. We report in Tab. 3 results of task-agnostic evaluations *(b1)*, which is our major concern in this work; methods without task-agnostic evaluations are absent in this table. As the protocol used in most current works, the results of task-aware evaluations on training split of the novel set *(a1)* are also reported in Tab. 4. Note that we report our reproduced results of UNO [13] on CIFAR10 using their officially debugged code (see Sec. C.1 in Appendix).

In Tab. 3, our proposed ComEx outperforms the current state of the arts with a clear margin, validating the efficacy of our proposed dividing-and-conquering solution. Specifically, CIFAR100-50 sees greater performance boost, indicating that current methods are easily misled by ambiguous boundaries between base and novel classes as discussed in Sec. 4.1. In contrast, our proposed ComEx works well on both sets of classes, benefiting from the comprehensive yet complementary abilities of two groups of experts. On the other hand, in Tab. 4 ComEx still yields competitive results with the traditional evaluation protocol only focusing on the training split of novel set. However, ComEx is inferior to MixUp [45] based methods [47, 48] which rely on synthesized strong negative samples during training. This strategy naturally benefits their NCD performance on the training split, which may arguably further improve ComEx likewise.

4.3. Qualitative Results

We provide in Fig. 4 the qualitative results of our proposed ComEx on testing split of CIFAR10. Figs. 4a and 4b show feature space (output of image encoder) visualizations before and after training. Since the image encoder is pre-trained on five base classes, at epoch 0 all samples are roughly grouped around these base classes, while at epoch 200 the ten classes are well separated in the feature space. In Figs. 4c and 4d we show output space visualizations by combining the outputs of each group of experts. In general, both batch- and class-wise experts produce separable outputs of the ten classes. Although Figs. 4c and 4d seem similar due to simplicity of the dataset, we can still observe that class-wise experts (Fig. 4d) yield clearer decision boundaries compared to batch-wise experts (Fig. 4c). This corresponds to our design purpose since each class-wise expert sees only a part of all classes, thus tends to produce (over-) confident outputs. This property facilitates recognizing images in most cases, yet may be harmful when facing hard samples as denoted in the red box in Fig. 4d. On the contrary, batch-wise experts are less-confident of their predictions due to exposed to more classes, thus can in turn tolerate hard samples by allowing local groupings as denoted in Fig. 4c. Again, our proposed ComEx leverages the comprehensive yet complementary abilities of both groups of experts, leading to superior GNCD performance.

5. Conclusion

We present in this paper ComEx, namely Compositional Experts, for the task of Generalized Novel Class Discovery (GNCD). Our goal is to generalize the recognition ability of traditional NCD to both base and novel sets. We show that current methods are far from achieving this goal, and propose to divide and conquer it with two groups of experts that characterize comprehensive yet complementary information of the data. We also introduce local-to-local aggregation to complement the widely-used global-to-local pseudo-labeling strategy, which considerably boosts the performance on recognizing novel classes. ComEx is evaluated on four GNCD benchmarks, demonstrating clear superiority against current state-of-the-art approaches. As future works, we plan to push the limits of GNCD to large-scale applications across different data regimes.

Acknowledgment

Our work was supported in part by the National Key R&D Program of China under Grant 2017YFE0104100; in part by the National Natural Science Foundation of China under Grants 62132016, 62171343, 62071361, and 62102293; in part by Key R&D Program of Shaanxi under Grant 2021ZDLGY01-03; and in part by the Fundamental Research Funds for the Central Universities ZDRC2102.

References

[1] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *Proc. ICLR*, 2020. 2, 3, 4, 5

[2] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Armand Joulin, Nicolas Ballas, and Michael Rabbat. Semi-supervised learning of visual features by non-parametrically predicting view assignments with support samples. In *Proc. ICCV*, 2021. 2, 5

[3] Yuval Atzmon, Felix Kreuk, Uri Shalit, and Gal Chechik. A causal view of compositional zero-shot recognition. In *Proc. NeurIPS*, 2020. 2

[4] David Berthelot, Nicholas Carlini, Ekin Dogus Cubuk, Alexey Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *Proc. ICLR*, 2020. 1, 2

[5] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Proc. NeurIPS*, 2019. 1, 2

[6] Jiarui Cai, Yizhou Wang, and Jenq-Neng Hwang. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *Proc. ICCV*, 2021. 4

[7] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proc. ECCV*, 2018. 2, 3, 6

[8] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Proc. NeurIPS*, 2020. 2, 3, 4, 5

[9] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *Proc. ECCV*, 2016. 1, 2

[10] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Proc. NeurIPS*, 2013. 3, 5

[11] Zhiyuan Dang, Cheng Deng, Xu Yang, Kun Wei, and Heng Huang. Nearest neighbor matching for deep clustering. In *Proc. CVPR*, 2021. 3

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. CVPR*, 2009. 1, 6

[13] Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *Proc. ICCV*, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[14] Kamran Ghasedi Dizaji, Amirhossein Herandi, Cheng Deng, Weidong Cai, and Heng Huang. Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. In *Proc. ICCV*, 2017. 3

[15] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *Proc. 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 *Proc. ICCV*, 2019. 1, 2, 6, 7, 8

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proc. CVPR*, 2020. 5

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016. 6

[19] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015. 4

[20] Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. In *Proc. ICLR*, 2018. 2, 7, 8

[21] Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. Multi-class classification without multi-class labels. In *Proc. ICLR*, 2019. 2, 7, 8

[22] Jiabo Huang, Qi Dong, Shaogang Gong, and Xiatian Zhu. Unsupervised deep learning by neighbourhood discovery. In *Proc. ICML*, 2019. 2, 3, 5

[23] Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proc. ICCV*, 2019. 6

[24] Xuihui Jia, Kai Han, Yukun Zhu, and Bradley Green. Joint representation learning and novel category discovery on single-and multi-modal data. In *Proc. ICCV*, 2021. 1, 2, 3, 6, 7, 8

[25] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv:1602.07332*, 2016. 1

[26] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[27] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 6

[28] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(3):453–465, 2013. 1, 2

[29] Junnan Li, Caiming Xiong, and Steven CH Hoi. Comatch: Semi-supervised learning with contrastive graph regularization. In *Proc. ICCV*, 2021. 2, 5

[30] Kai Li, Martin Renqiang Min, and Yun Fu. Rethinking zero-shot learning: A conditional visual classification perspective. In *Proc. ICCV*, 2019. 2

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. ECCV*, 2014. 1

[32] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proc. BSMSP*, 1967. 7

[33] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han

Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Proc. NeurIPS*, 2020. 2

[34] Laurens Van Der Maaten. Accelerating t-sne using tree-based algorithms. *J. Mach. Learn. Res.*, 15(1):3221–3245, 2014. 8

[35] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *Proc. ECCV*, 2020. 3, 5

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Proc. NeurIPS*, 2017. 5

[37] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(9):2251–2265, 2018. 1, 2

[38] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *Proc. ICML*, 2016. 3

[39] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. In *Proc. NeurIPS*, 2020. 1, 2

[40] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. In *Proc. NeurIPS*, 2020. 2

[41] Jianwei Yang, Devi Parikh, and Dhruv Batra. Joint unsupervised learning of deep representations and image clusters. In *Proc. CVPR*, 2016. 3

[42] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv:2110.11334*, 2021. 4

[43] Muli Yang, Cheng Deng, Junchi Yan, Xianglong Liu, and Dacheng Tao. Learning unseen concepts via hierarchical decomposition and composition. In *Proc. CVPR*, 2020. 2

[44] Xu Yang, Cheng Deng, Kun Wei, Junchi Yan, and Wei Liu. Adversarial learning for robust deep clustering. In *Proc. NeurIPS*, 2020. 3

[45] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proc. ICLR*, 2018. 3, 8

[46] Bingchen Zhao and Kai Han. Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. *arXiv:2107.03358*, 2021. 1, 2, 3, 6, 7, 8

[47] Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. Neighborhood contrastive learning for novel class discovery. In *Proc. CVPR*, 2021. 1, 2, 3, 5, 6, 7, 8

[48] Zhun Zhong, Linchao Zhu, Zhiming Luo, Shaozi Li, Yi Yang, and Nicu Sebe. Openmix: Reviving known knowledge for discovering novel visual categories in an open world. In *Proc. CVPR*, 2021. 1, 2, 3, 5, 6, 7, 8

[49] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *Proc. ICCV*, 2019. 3, 5