Contrastive Mean-Shift Learning for Generalized Category Discovery

Sua Choi Dahyun Kang Minsu Cho
Pohang University of Science and Technology (POSTECH), South Korea

Abstract

We address the problem of generalized category discovery (GCD) that aims to partition a partially labeled collection of images; only a small part of the collection is labeled and the total number of target classes is unknown. To address this generalized image clustering problem, we revisit the mean-shift algorithm, i.e.*, a classic, powerful technique for mode seeking, and incorporate it into a contrastive learning framework. The proposed method, dubbed Contrastive Mean-Shift (CMS) learning, trains an image encoder to produce representations with better clustering properties by an iterative process of mean shift and contrastive update. Experiments demonstrate that our method, both in settings with and without the total number of clusters being known, achieves state-of-the-art performance on six public GCD benchmarks without bells and whistles.*

1. Introduction

Clustering is one of the most fundamental problems in unsupervised learning, which aims to partition instances of a data collection into different groups [1, 10, 19, 23]. Unlike the classification problem, it does not assume either predefined target classes or labeled instances in its standard setup. However, in a practical scenario, some data instances may be labeled for a subset of target classes so that we can leverage them to cluster all the data instances while also discovering the remaining unknown classes. The goal of Generalized Category Discovery (GCD) [25] is to tackle such a semi-supervised image clustering problem given a small amount of incomplete supervision.

Viewing GCD as a transductive learning problem for semi-supervised clustering, we revisit the mean shift [5, 7, 12], *i.e.*, a classic, powerful technique for mode seeking and clustering analysis. The mean-shift algorithm assigns each data point a corresponding mode through iterative shifts by kernel-weighted aggregation of neighboring points so as to cluster the data points according to their modes; this process is non-parametric and does not require any information about the target clusters, *e.g.*, the number of clusters.

By incorporating the mean shift into a contrastive learning framework [4, 15, 32], we introduce *contrastive mean-shift learning* that induces an embedding space with better clustering properties for GCD.

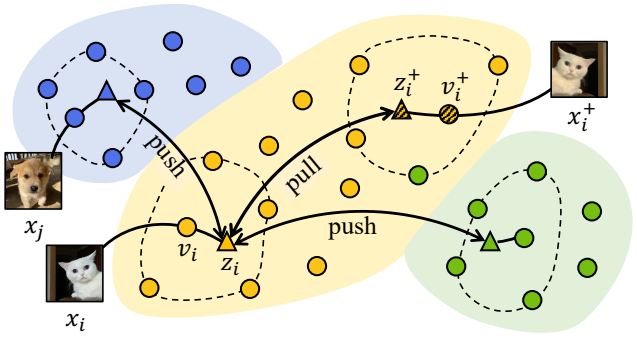

Figure 1. **Contrastive Mean-Shift (CMS) learning.** By integrating mean shift [7, 12] into contrastive learning [4, 32], the proposed method learns an embedding space such that the mean-shifted embeddings of identical images x_i and x_i^+ draw together and those of distinct images x_i and x_j push apart from each other.

Prior arts for GCD [6, 21, 25, 28, 30] typically employ k-means clustering [1, 19] in validation and testing where the ground-truth number of target classes K is often required for stable performance, which is undesirable in practical scenarios. In contrast, our method jointly estimates K during training so that it achieves robust performance without using the ground-truth K in clustering.

The proposed method is extensively evaluated on the six public GCD datasets [13, 16, 17, 20, 24, 26], including coarse-grained and fine-grained classification problems, and achieves the state-of-the-art performance on the public GCD benchmarks [25, 30]. Notably, even when the ground-truth number of target classes K is not used, it shows comparable performance to the state-of-the-art methods that exploit the ground-truth K.

Our contribution can be summarized as follows:

- We introduce contrastive mean-shift learning for GCD by incorporating the mean-shift algorithm in a contrastive learning framework.

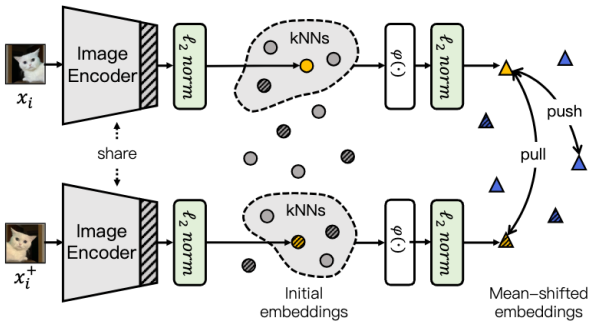

Figure 2. **Contrastive Mean-Shift Learning.** Given a collection of images, each initial image embedding $\boldsymbol{v}_i$ from an image encoder takes a single step of mean shift to be $\boldsymbol{z}_i$ by aggregating its k nearest neighbors with a weight kernel $\varphi(\cdot)$. The encoder network is then updated by contrastive learning with the mean-shifted embeddings, which draws a mean-shifted embedding of image x_i and that of its augmented image x_i^+ closer and pushes those of distinct images apart from each other.

- The proposed method jointly estimates the number of target classes in training and thus achieves robust discovery without using the ground-truth number of target classes.
- Extensive experiments and analyses demonstrate that our method outperforms the state-of-the-art methods on several standard GCD benchmarks.

2. Preliminary: mean-shift algorithm

Given a collection of data points $\mathcal{V}$ in a feature space, the weighted mean $m(\boldsymbol{v}_i)$ of each data point $\boldsymbol{v}_i$ is calculated over its neighborhood $\mathcal{N}(\boldsymbol{v}_i) \subseteq \mathcal{V}$ as:

$$m(\boldsymbol{v}_i) = \frac{\sum_{\boldsymbol{v}_j \in \mathcal{N}(\boldsymbol{v}_i)} \varphi(\boldsymbol{v}_j - \boldsymbol{v}_i)\boldsymbol{v}_j}{\sum_{\boldsymbol{v}_j \in \mathcal{N}(\boldsymbol{v}_i)} \varphi(\boldsymbol{v}_j - \boldsymbol{v}_i)}, \tag{1}$$

where a kernel function $\varphi(\cdot)$ determines weights for neighbors in estimating the mean. The mode of $\boldsymbol{v}_i$ is sought by iteratively shifting to its weighted mean until convergence. The set of data points that converge to the same mode defines the basin of attraction of that mode, and this naturally relates to clustering: the points in the same basin of attraction are associated with the same cluster [7].

The mean shift is characterized by the set of neighbors $\mathcal{N}(\boldsymbol{v}_i)$ and the kernel function $\varphi(\cdot)$. In typical setups [5, 7, 8, 29], $\mathcal{N}(\boldsymbol{v}_i)$ is defined by a certain radius and $\varphi(\cdot)$ is set to a uniform, Gaussian, or Epanechnikov kernel [22].

3. Contrastive Mean-Shift

We propose *contrastive mean-shift learning* for GCD by integrating the mean shift in the contrastive learning framework. Figure 2 illustrates the overall procedure.

Given a collection of images with partial labels, we obtain initial image embeddings from a self-supervised encoder network [3], perform a single-step mean shift on each of them using its k nearest neighbors (kNNs) (Sec. 3.1), and then update the last layer of the encoder through contrastive learning [4, 15, 32] across the mean-shifted embeddings (Sec. 3.2). After each epoch of training, the number of classes K is estimated by agglomerative clustering [27] and used to measure the clustering accuracy of the snap-shot model on the validation set (Sec. 3.3). This update procedure is performed for a sufficient number of epochs, and the best model is selected according to the validation accuracy.

After training the encoder, we apply multi-step mean shifts on the final embedding space and the final cluster assignment is performed using the number of clusters K estimated in training (Sec. 3.4).

3.1. Mean-shifted embedding

Given a collection of images $\mathcal{X} = \{x_1, x_2, \cdots, x_N\}$, the images are fed to an image encoder f to generate the corresponding set of d-dimensional $l2$-normalized embeddings: $\mathcal{V} = \{\boldsymbol{v}_i\}_{i=1}^N$, where $\boldsymbol{v}_i = f(x_i)$. To obtain discriminative initial embeddings without supervision, we use the self-supervised pre-trained encoder, DINO [3]; our method is not restricted to a specific encoder.

The *mean-shifted embedding* $\boldsymbol{z}_i$ is obtained from the initial embedding $\boldsymbol{v}_i$ using a single-step mean shift similar to Eq. (1). The conventional mean shift typically defines the neighborhood for each data point based on a distance, *i.e.*, radius. We find that the number of neighbors within a fixed radius varies significantly during the update of the encoder, causing the training phase to be unstable. To address the issue, we replace the distance-based NNs with kNNs, which greatly improves the stability and is also suitable for parallel computation with GPUs. The neighborhood $\mathcal{N}(\boldsymbol{v}_i)$ is thus defined with input $\boldsymbol{v}_i$ and its kNNs:

$$\mathcal{N}(\boldsymbol{v}_i) = \{\boldsymbol{v}_i\} \cup \text{argmax}^k_{\boldsymbol{v}_j \in \mathcal{V}} \boldsymbol{v}_i \cdot \boldsymbol{v}_j, \tag{2}$$

where $\text{argmax}^k_{s \in \mathcal{S}}(\cdot)$ returns a subset of the top-k items that maximizes a target function.

Along with the neighborhood, we design the weight kernel $\varphi(\cdot)$ to put a higher weight on the center, *i.e.*, the query position $\boldsymbol{v}_i$, compared to its kNNs in aggregation:

$$\varphi(\boldsymbol{v}) = \begin{cases} 1-\alpha & \text{if } ||\boldsymbol{v}|| = 0 \\ \frac{\alpha}{k} & \text{otherwise,} \end{cases} \tag{3}$$

where α denotes a scaling hyperparameter ($\alpha = 0.5$ in our experiment). This kernel can be interpreted as a simple approximation of a Gaussian kernel with adaptive bandwidth.

The mean-shifted embedding $\boldsymbol{z}_i$ is obtained by aggregating the neighbor embeddings with the kernel and then $l2$-normalizing it, ensuring that the shifted embedding remains on a unit hypersphere. We use these mean-shifted embeddings in contrastive learning to update the encoder, which is described next.

Method	CIFAR100			ImageNet100			CUB			Stanford Cars			FGVC Aircraft			Herbarium 19		
	All	Old	Novel	All	Old	Novel	All	Old	Novel	All	Old	Novel	All	Old	Novel	All	Old	Novel
(a) Clustering with the ground-truth number of classes K given																		
Agglomerative [27]†	56.9	56.6	57.5	73.1	77.9	70.6	37.0	36.2	37.3	12.5	14.1	11.7	15.5	12.9	16.9	14.4	14.6	14.4
RankStats+ [14]	58.2	77.6	19.3	37.1	61.6	24.8	33.3	51.6	24.2	28.3	61.8	12.1	26.9	36.4	22.2	27.9	55.8	12.8
UNO+ [11]	69.5	80.6	47.2	70.3	95.0	57.9	35.1	49.0	28.1	35.5	70.5	18.6	40.3	56.4	32.2	28.3	53.7	14.7
ORCA [2]	69.0	77.4	52.0	73.5	92.6	63.9	35.3	45.6	30.2	23.5	50.1	10.7	22.0	31.8	17.1	20.9	30.9	15.5
GCD [25]	73.0	76.2	66.5	74.1	89.8	66.3	51.3	56.6	48.7	39.0	57.6	29.9	45.0	41.1	46.9	35.4	51.0	27.0
DCCL [21]	75.3	76.8	70.2	80.5	90.5	76.2	63.5	60.8	**64.9**	43.1	55.7	36.2	-	-	-	-	-	-
PromptCAL [30]	81.2	84.2	75.3	83.1	92.7	78.3	62.9	64.4	62.1	50.2	70.1	40.6	52.2	52.2	**52.3**	37.0	52.0	28.9
GPC [31]	77.9	85.0	63.0	76.9	94.3	71.0	55.4	58.2	53.1	42.8	59.2	32.8	46.3	42.5	47.9	-	-	-
SimGCD [28]	80.1	81.2	**77.8**	83.0	93.1	77.9	60.3	65.6	57.7	53.8	71.9	45.0	54.2	59.1	51.8	**44.0**	**58.0**	**36.4**
PIM [6]	78.3	84.2	66.5	83.1	95.3	77.0	62.7	75.7	56.2	43.1	66.9	31.6	-	-	-	42.3	56.1	34.8
Ours	**82.3**	**85.7**	75.5	**84.7**	**95.6**	**79.2**	**68.2**	**76.5**	64.0	**56.9**	**76.1**	**47.6**	**56.0**	**63.4**	**52.3**	36.4	54.9	26.4
(b) Clustering without the ground-truth number of classes K given																		
Agglomerative [27]†	56.9	56.6	57.5	72.2	77.8	69.4	35.7	33.3	36.9	10.8	10.6	10.9	14.1	10.3	16.0	13.9	13.6	14.1
GCD [25]	70.8	77.6	57.0	77.9	91.1	71.3	51.1	56.4	48.4	39.1	58.6	29.7	-	-	-	37.2	51.7	29.4
GPC [31]	75.4	**84.6**	60.1	75.3	93.4	66.7	52.0	55.5	47.5	38.2	58.9	27.4	43.3	40.7	44.8	36.5	51.7	27.9
PIM [6]	75.6	81.6	63.6	**83.0**	95.3	**76.9**	62.0	**75.7**	55.1	42.4	65.3	31.3	-	-	-	**42.0**	55.5	**34.7**
Ours	**79.6**	83.2	**72.3**	81.3	**95.6**	74.2	**64.4**	68.2	**62.4**	**51.7**	**68.9**	**43.4**	**55.2**	**60.6**	**52.4**	37.4	**56.5**	27.1

Table 1. Comparison with the state of the arts on GCD, evaluated *with* or *without* the GT K for clustering. † denotes reproduced results.

	training	inference		CIFAR100			ImageNet100			CUB			Stanford Cars		
	CMS	SSK	IMS	All	Old	Novel	All	Old	Novel	All	Old	Novel	All	Old	Novel
(1)		✓		71.5	77.3	60.1	74.1	89.8	66.3	51.2	49.2	52.2	37.9	57.8	28.3
(2)			✓	71.6	77.3	60.0	80.3	91.7	74.6	58.7	62.0	57.1	40.8	54.5	34.2
(3)	✓	✓		81.1	85.6	72.1	83.4	**95.8**	77.2	66.7	75.3	62.5	54.5	76.4	43.9
(4)	✓		⋄ (1-step)	80.1	**86.0**	68.4	84.1	95.6	78.3	**68.2**	76.4	**64.1**	56.1	74.6	47.1
(5)	✓		✓	**82.3**	85.7	**75.5**	**84.7**	95.6	**79.2**	**68.2**	**76.5**	64.0	**56.9**	**76.1**	**47.6**

Table 2. Effectiveness of each component of our method. SSK denotes semi-supervised k-means clustering and IMS iterative mean-shift.

3.2. Contrastive mean-shift learning

The objective of contrastive mean-shift learning is to encourage the model to improve its clustering properties in the mean-shifted embedding space, bringing closer the mean-shifted embeddings of an identical image while pushing apart those of distinct images. The learning objective consists of two types of terms: (1) the unsupervised contrastive mean-shift loss $\mathcal{L}_{\text{CMS}}$ for all images $\mathcal{D}_{\text{L}} \cup \mathcal{D}_{\text{UL}}$ and (2) the supervised contrastive loss $\mathcal{L}_{\text{SC}}$ for the labeled set $\mathcal{D}_{\text{L}}$.

We apply random image augmentation with cropping, flipping and color jittering [25] to all images in a batch and create positive pairs, x_i and its augmented version x_i^+ while considering pairs of two distinct images, x_i and x_j, as negative pairs. The unsupervised contrastive mean-shift loss is designed to decrease the distance between the mean-shifted embeddings of the positive pair and increase the distance between those of the negative pair. The individual loss term for the mean-shifted embedding $\boldsymbol{z}_i$ is formulated as:

$$\mathcal{L}_{\text{CMS}}^{(i)} = -\log \frac{\exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_i^+ / \tau_u)}{\sum_{j \neq i} \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_j / \tau_u)}, \tag{4}$$

where τ_u is a hyperparameter for adjusting the temperature.

Similarly, the supervised contrastive learning loss [15, 25] is formed with the labeled images only, which decreases the distance between the features of the same class and increases the distance between the others according to the given ground-truth class labels:

$$\mathcal{L}_{\text{SC}}^{(i)} = -\frac{1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\exp(\boldsymbol{v}_i \cdot \boldsymbol{v}_p / \tau_s)}{\sum_{j \notin \mathcal{P}(i)} \exp(\boldsymbol{v}_i \cdot \boldsymbol{v}_j / \tau_s)}, \tag{5}$$

where $\mathcal{P}(i)$ represents the set of image indices for the same class with image x_i in a batch.

Denoting by $\mathcal{B}$ the set of image indices in a batch and by $\mathcal{B}_{\text{L}}$ its subset for labeled images, the overall learning objective combines the two types of individual losses:

$$\mathcal{L} = \lambda \frac{1}{|\mathcal{B}_{\text{L}}|} \sum_{i \in \mathcal{B}_{\text{L}}} \mathcal{L}_{\text{SC}}^{(i)} + (1-\lambda) \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \mathcal{L}_{\text{CMS}}^{(i)}, \tag{6}$$

where λ represents the hyperparameter balancing between two types of losses.

3.3. Estimating the number of clusters

During training, we also estimate the number of clusters K by measuring the clustering accuracy on the validation set $\mathcal{D}_{\text{V}}$. For actual clustering, we use the agglomerative clustering algorithm with the ward linkage criterion [27], which iteratively merges the closest pair of clusters until it reaches a certain distance threshold or a target number of clusters. At the end of each training epoch, we apply the

algorithm to $\mathcal{D}_{\text{V}}$ and obtain clustering results for different numbers of clusters. Among them, the highest clustering accuracy, which is measured using the labeled images in $\mathcal{D}_{\text{V}}$, is recorded as the validation performance at the epoch and the corresponding number of clusters is determined as the estimated number of clusters K at the epoch. Once the training process is over, the snapshot model at the epoch with the best validation performance is selected as the final model, and the estimated K at the epoch is determined as the final estimation of the number of clusters. This approach allows us to avoid accessing the ground-truth number of classes during both training and validation, in contrast to previous work [6, 21, 25, 28, 30].

3.4. Final clustering inference

The encoder learned by contrastive mean-shift learning is used for the final cluster assignment; we extract embeddings of images using the encoder and partition them into K clusters using agglomerative clustering with the estimated K. To improve the clustering property of the embeddings, we perform multi-step mean shift on the embeddings before agglomerative clustering. Starting from the initial embeddings $\mathcal{V}^{(0)}$ from the learned encoder, we update them to t-step mean-shifted embeddings $\mathcal{V}^{(t)}$ until the clustering accuracy on $\mathcal{D}_{\text{L}}$ does not increase for two consecutive iterations. The final cluster assignment is obtained by agglomerative clustering on the multi-step mean-shifted embeddings.

4. Experiments

4.1. Experimental setup

Training details. We follow the standard training datasets, evaluation benchmark, and protocols of the existing work on GCD [21, 25, 30]. We use the pre-trained DINO ViT-B/16 [3, 9] with a projection head as our image encoder. We used a single RTX-3090 for all experiments.

Evaluation. The accuracy is measured by matching the assignments with ground-truth labels by the Hungarian optimal matching [18], based on the number of intersected instances between each pair of classes. The unpaired classes are considered incorrect predictions, while the instances of the most dominant class within each ground-truth cluster are considered correct when calculating the accuracy. The accuracy is reported on "All" unlabeled data as well as the accuracy on those of known and unknown classes, denoted by "Old" and "Novel" in tables, respectively.

4.2. Main results

Evaluation on GCD. Table 1 presents a comparison on the GCD setup in both coarse-grained and fine-grained benchmarks with or without the ground-truth (GT) number of classes K. In Table 1 (a), we compare our method with the state-of-the-art methods, all evaluated with GT K. Note that the GT K is only used for evaluation purpose in our case, and not for model selection during training. The other state-of-the-art methods adopt semi-supervised k-means clustering, where the K centroids are initialized by the labeled data with the GT K. Our method achieves state-of-the-art performance on five out of six datasets.

In Table 1 (b), we present the comparison of ours and the state of the arts on the same setup with Table 1 (a) but without having the GT number of classes K known for clustering. For Vaze *et al.* [25], we borrow the results from the work of PIM [6]. *Our method shows outstanding performance in most scenarios even though it does not access to the GT K in both training and testing.* Our method is even superior to the state-of-the-art methods measured with the known value of K on CUB and FGVC Aircraft. The results show that our K-estimation process incorporated in the training phase performs effectively with no significant performance drop compared to the known-K counterparts.

Effect of each proposed component. Table 2 shows the ablation of CMS learning (Sec. 3.2) and Iterative Mean Shift (IMS, Sec. 3.4) for final clustering inteference. For training, we examine the effect of the embedding without mean shift, *i.e.*, equivalent to the embedding in Vaze *et al.* [25]. At inference, semi-supervised k-means clustering (SSK) [25], single-step mean shift, and IMS are compared. Comparing (1) *vs* (3) and (2) *vs* (5), we observe that CMS learning boosts performance significantly. After training, IMS brings additional gains at inference when comparing (1) *vs* (2) and (3) *vs* (5), plus recursive iterations: (4) *vs* (5). The final model (5) outperforms others with the combined gain of each proposed component.

5. Conclusion

We have proposed to revisit the mean-shift algorithm and incorporated it with contrastive representation learning for generalized category discovery. While the previous work on GCD often exploits the ground-truth number of classes for clustering, we avoid using the oracle information and instead estimate the number of clusters using agglomerative clustering. Our method achieves state-of-the-art performance on public GCD benchmarks without bells and whistles. We believe that the proposed contrastive mean-shift learning will benefit representation learning for other diverse tasks beyond category discovery.

Acknowledgments. This work was supported by the NRF grant (RS-2021-NR059830 (50%)) and the IITP grants (RS-2022-II220113: Developing a Sustainable Collaborative Multi-modal Lifelong Learning Framework (25%), RS-2024-00457882: National AI Research Lab Project (20%), RS-2019-II191906: AI Graduate School Program at POSTECH (5%)) funded by Ministry of Science and ICT, Korea.

References

[1] David Arthur and Sergei Vassilvitskii. K-means++ the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035, 2007. 1

[2] Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. In *Proc. International Conference on Learning Representations (ICLR)*, 2022. 3

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 2, 4

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proc. International Conference on Machine Learning (ICML)*, 2020. 1, 2

[5] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 1995. 1, 2

[6] Florent Chiaroni, Jose Dolz, Ziko Imtiaz Masud, Amar Mitiche, and Ismail Ben Ayed. Parametric information maximization for generalized category discovery. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 1729–1739, 2023. 1, 3, 4

[7] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2002. 1, 2

[8] Dorin Comaniciu, Visvanathan Ramesh, and Peter Meer. Real-time tracking of non-rigid objects using mean shift. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2000. 2

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. International Conference on Learning Representations (ICLR)*, 2021. 4

[10] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, 1996. 1

[11] Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 3

[12] Keinosuke Fukunaga and Larry Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on information theory*, 1975. 1

[13] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *Proc. International Conference on Learning Representations (ICLR)*, 2019. 1

[14] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *Proc. International Conference on Learning Representations (ICLR)*, 2020. 3

[15] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2, 3

[16] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 1

[17] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, 2009. 1

[18] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 4

[19] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Oakland, CA, USA, 1967. 1

[20] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 1

[21] Nan Pu, Zhun Zhong, and Nicu Sebe. Dynamic conceptional contrastive learning for generalized category discovery. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7579–7588, 2023. 1, 3, 4

[22] David W Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015. 2

[23] Robin Sibson. Slink: an optimally efficient algorithm for the single-link cluster method. *The computer journal*, 1973. 1

[24] Kiat Chuan Tan, Yulong Liu, Barbara Ambrose, Melissa Tulig, and Serge Belongie. The herbarium challenge 2019 dataset. *arXiv preprint arXiv:1906.05372*, 2019. 1

[25] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 3, 4

[26] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical report, 2011. 1

[27] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963. 2, 3

[28] Xin Wen, Bingchen Zhao, and Xiaojuan Qi. Parametric classification for generalized category discovery: A baseline study. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 16590–16600, 2023. 1, 3, 4

[29] Kuo-Lung Wu and Miin-Shen Yang. Mean shift-based clustering. *Pattern Recognition*, 2007. 2

[30] Sheng Zhang, Salman Khan, Zhiqiang Shen, Muzammal Naseer, Guangyi Chen, and Fahad Shahbaz Khan. Promptcal: Contrastive affinity learning via auxiliary prompts for

generalized novel category discovery. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3479–3488, 2023. 1, 3, 4

[31] Bingchen Zhao, Xin Wen, and Kai Han. Learning semi-supervised gaussian mixture models for generalized category discovery. *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2023. 3

[32] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019. 1, 2